# Online Query Refinement on Information Retrieval Systems: A Process Model of Searcher/System Interactions

*Hsinchun Chen*[1], MIS Department, University of Arizona
*Vasant Dhar*[2], IS Department, New York University

Abstract

This article reports findings of empirical research that investigated information searchers' online query refinement process. Prior studies have recognized the information specialists' role in helping searchers articulate and refine queries. Using a semantic network and a Problem Behavior Graph to represent the online search our study revealed that searchers also refined their own queries in an online task environment. The information retrieval system played a passive role in assisting online query refinement, which was, however, one that confirmed Taylor's four-level query formulation model. Based on our empirical findings, we proposed using process model to facilitate and improve query refinement in an online environment. We believe incorporating this model into retrieval systems can result in the design of more "intelligent" and useful information reuieval systems.

## Introduction

Electronic text-based information storage and retrieval systems in the form of online catalogs, online bibliographic databases. and videotex that can store huge amounts of data and allow access via a terminal or a television set are changing the way we gather, process. and retrieve information. These systems provide a wide variety of information and services, ranging from daily updates of foreign and national news, movie reviews, law cases, and financial data on companies to journal articles, books, trademarks, and statistics. While archival information sources such as libraries are becoming increasingly computerized, access to such information is often difficult, due in large pan to the indeterminate nature of the process by which documents are indexed and the latitude searchers have in expressing a query. For inexperienced searchers. the problem of finding documents that are relevant to a query can be difficult for three reasons:

1. it can require a significant amount of knowledge of the subject area in which information is sought,
2. it requires knowledge about the functionality of the information storage and retrieval system, and
3. it requires knowledge about the classification scheme employed in the information storage and retrieval system.

Searchers generally have limited knowledge about the classification scheme and the retrieval system. Since the purpose of the search itself is often to acquire knowledge about the subject area. searchers may not be clear about the subject area for which answers are being sought Searchers may have only a felt or conscious need which requires to be formalized and articulated. A human information specialist such as a reference librarian often assumes an active role in helping searchers refine and articulate their queries.

The focus of our research was to identify empirically how searchers refined their queries when using an online catalog. We compared the results with prior studies in which librarians' assistance was present. Based on our empirical findings. we proposed a process model for facilitating online query refinement.

[1]hchen@mis.arizona.edu, MIS, University of Arizona, Tucson. AZ 85721, USA
[2]vdh@vxl.pba.nyu.edu. IS, NYU. NY. NY 10003, USA

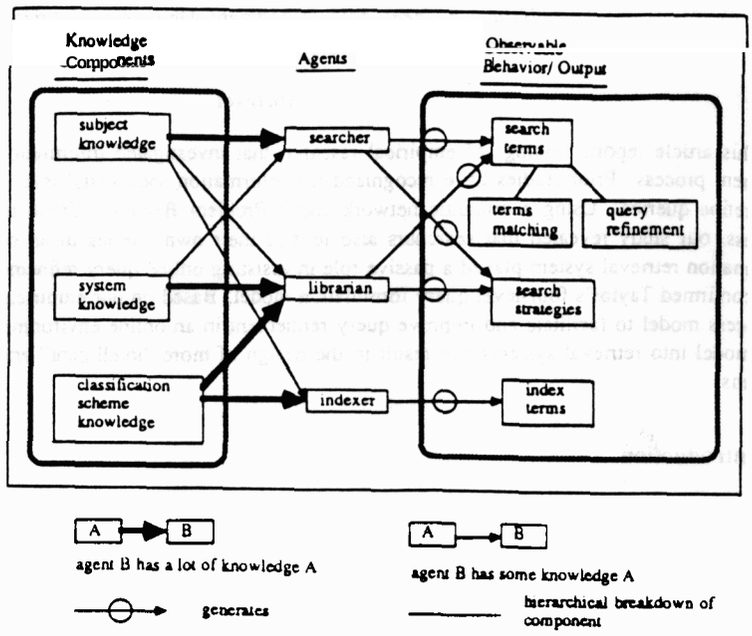


Figure A framework for information storage and retrieval

In Section 2, we present an information retrieval framework as the context for our research. We then describe our research design and method in Section 3. In Section 4, we discuss how searchers refined their queries in an online environment based on Taylor's theory of query formulation. In Section 5, we present a process model that aimed at facilitating online query refinement. We draw conclusion concerning our research in the final section.

## 2 Information Retrieval Framework

In this section, we present the information retrieval framework we developed. This framework, as shown in Figure 1, examines the human agents involved in the online information storage and retrieval environment. the types of knowledge these agents possess, and the unique characteristics of their indexing and search behavior

### 2.1 Knowledge Components

Three types of knowledge are required for the information storage and retrieval process. First, classification scheme knowledge is used for indexing documents. Second, subject area knowledge is required for expressing a query. Lastly, system knowledge helps a user perform a fruitful and effective search in an online retrieval system. These three knowledge components are presented on the left hand side of Figure 1.

## 2.2 Agents

The three types of knowledge are typically distributed among three different parties to whom we refer as "human agents." These agents include *indexers*, who classify the documents based on some pre-determined classification scheme; *searchers*, who express their queries in their own terms; and *reference librarians*, who serve as intermediaries between searchers and retrieval systems.

Searchers generally do not have classification scheme knowledge and their knowledge of the subject area and the system functionality varies widely. Indexers generally have a great deal of classification scheme knowledge and some subject knowledge. They are, however, not concerned about the functionality of retrieval systems. Librarians must have all three kinds of knowledge, although they generally know more about the classification scheme and system than about various subjects. The relationships between the knowledge components and the agents are represented as links in Figure 1.

## 2.3 Indexing and Search Behavior

Indexing uncertainty and search uncertainty are the primary sources of information retrieval problems.

### 2.3.1 Indexing Uncertainty

The process of indexing is partly indeterminate. Evidence suggests that different indexers, all of whom are well trained in an indexing scheme, might assign different index terms to a given document (see the box labeled index terms in Figure 1). It has also been observed that an indexer may use different terms for the same document at different times [15] [24].

### 2.3.2 Search uncertainty

Search uncertainty refers to the latitude searchers have in adopting search strategies and choosing search terms (see the boxes on the right side of Figure 1) during the information reuieval process.

1. Search Strategies:

   *Search strategy* is often used to describe the plan or approach to the whole search. In a card catalog study, two strategies for searching have been identified: a "self-reliant" style where searchers generate their own search terms and a "caialog-oriented" style where searchers use the *terms* found in the card catalog [26]. Another study classifies the search strategy in terms of the critical decision points faced during the online search. Two *types of* decision points occur during the search: a decision to react to unfavorable results and a decision to revise search logic [16]. Bourne identifies two search strategies. In the "building-block" approach. one enters various terms as separate search statements. After the search results a n derived, one combines all search statements into a single final statement using the Boolean operator. AND. This strategy contrasts with the "pearl-growing" strategy. in which one starts by searching on a few specific terms to retrieve initial citations. These citations arc then examined carefully for new candidau search terms to be added in the subsequent searches [18].

2. Search Terms:

   A high degree of uncertainty with regard to search terms has been observed. Searchen tend to use different search terms for the same information sought. Studies have revealed *that,* on average, the probability of any two people using the same term to describe an object is about 10 percent [13] [12]. This fundamental

property of language limits the success of various design methodologies for controlled vocabulary-driven interaction [12].

Due to the uncertainty in choosing the index and the search terms, generating an exact march between the searcher's terms and those of the indexer becomes difficult. Bates [2] argues that for a successful match, the searcher must somehow generate as much "variety" (in the cybernetic sense, as defined by [1]) in the search as is produced by the indexers in their indexing.

While indexers use the rule of specificity for indexing, searchers tend to approach a search by specify broader terms first. There may be several reasons for this. One hypothesis is that searchers often do not have "queries," but what Belkin calls an "anomalous state of knowledge" [4]. Searchers often expect to refine this anomalous state into a query *through* an interactive process. The organization of a catalog or a system does not always facilitate this type of query refinement. however. In contrast, reference librarians appear to be particularly adept at performing this function.

Taylor suggests that a searcher's queries start from an actual but unexpressed need (visceral need). The visceral need is refined to a conscious description of the need (conscious need). This need is finally formalized as a statement (formalized need). However, the actual query presented to the information system may be compromised by the searcher's expectation of the system (compromised need) [28]. Based on Taylor's model. a similar model for describing query refinement during the pre-search interview between reference librarians and online searchers was developed by Markey [17]. We also used Taylor's theory to model the online query refinement process. Details are discussed in Section 4.

The importance of query refinement during the information retrieval process and the reference librarian's role in assisting this process are well recognized in earlier research Nevertheless, prior studies did not investigate the functionality of the retrieval systems in assisting query refinement. In our research, we investigated online query refinement process in detail, and our findings were used to develop a process model for aiding online information retrieval.

## 3 Research Design

A field study was conducted in 1988 at New York University. Data collection techniques including think-aloud protocols, tape-recordings, interviews. and questionnaires were used. By studying the interactions between information searchers and an online catalog, we were able to identify the query refinement proccss whch occurs The online catalog system we studied, Bobcat. listed over 600,000 catalog records including all new materials acquired after 1973 and many older items previously listed in the card catalog. Joumals were not listed. The system provided seven search options: tide search. author search. combination of author and title search, subject search, call number search, keyword search, and Boolean search. These options are considered standard in most online catalog systems.

Thirty business school students ranging from Ph.D. candidates to freshmen participated in the study. These subjects were asked to perform a search for documents within a subject area of their own choosing. In general the most frequently chosen option was subject search, followed by keyword search using index term (one word only). Before beginning their searches, subjects were requested to explain what they were looking for. They were also asked to think aloud during their searches. Their verbal protocols were tape-recorded, and the interactions between the searchers and the system were logged. The interactions lasted between 5 and 40 minutes. After the interaction, subjects were requested to state their queries again. A few follow-up questions pertaining to the search process and the problems encountered during the search also were asked.

Our study was based on information processing theory [19]. The technique we used to analyze the collected

data was ***protocol analysis.*** Cognitive Psychology comm

## 4 Levels of Query Refi

Based on the representation describe the online query ref the differences between the such help.

### 4.1 Representation of S

The representation scheme *network* and a ***Problem*** Beh Intelligence. The semantic r on the other hand. depicts th

#### 4.1.1 Semantic Network

The concept of a semantic n encoding the meaning of wo represent objects, things, cor been used in different applic also have been applied to th for retrieval systems. have t

Knowledge Of a subject where links are of two type general and specific terms ( Subject Headings (LCSH) h a portion of the semantic ne

#### 4.1.2 Problem Behavior

In order to make the analysi process in terms of a Probl in a time sequence, from a the needs). This detailed re interaction logs and verbal ***operator elements*** [19] [31 task The ***operator element*** state of knowledge as outp PBG, which was grounded to study the human problem analysis [5], mathematical p useful to represent the onlin

data was *protocol analysis*, a qualitative analysis technique frequently used in the Artificial Intelligence and Cognitive Psychology communities [11].

## 4 Levels of Query Refinement in Online Search

Based on the representation we developed for capturing the online query refinement process, we were able to describe the online query refinement process in terms of Taylor's theory. We conclude this section by discussing the differences between the way librarians assist in refining queries and the way information systems provide such help.

### 4.1 Representation of Search Process

The representation scheme we used for the analysis of the query refinement process is based on a ***semantic network*** and a ***Problem Behavior*** Graph (PBG). Both representations are widely used in the area of Artificial Intelligence. The semantic network is used to represent the semantic contents of searchers' queries. The PBG, on the other hand, depicts the flow of the online information retrieval process.

#### 4.1.1 Semantic Network

The concept of a semantic network was first introduced by Quillian [21] as a general association mechanism for encoding the meaning of words. A semantic network represents knowledge by means of nodes and links: nodes represent objects, things, concepts, facts, etc.; links represent relationships among them. Semantic networks have been used in different applications, mainly in the area of natural language processing [23] [33] [6]. Recently, they also have been applied to the design of information retrieval systems [22] [9]. In particular, the online thesauri for retrieval systems, have been represented using a semantic network structure [10] [25].

Knowledge of a subject area can be captured and represented by a large semantic network of terms (concepts) where links are of two types: relations between non-index and index terms (USE links) and relations between general and specific terms (NT/BT links). These links exist in most thesauri, including the Library of Congress Subject Headings (LCSH) handbook (the classification scheme of the online catalog we studied). **Figure** 2 shows a portion of the semantic network corresponding to the LCSH classification scheme.

#### 4.1.2 Problem Behavior Graph

In order to make the analysis of the online query refinement process more meaningful, we summarize the search process in terms of a Problem Behavior Graph (PBG). This representation describes problem-solving activities in a time sequence, from an initial state (a vague description of needs) to a goal state (a solution that satisfies the needs). This detailed representation of the problem-solving process is derived by first splitting up the user's interaction logs and verbal protocols into their ***semantic elements***, which consists of ***knowledge elements*** and ***operator elements*** [19] [31]. The ***knowledge elements*** specify the kinds of knowledge the subject has about the task. The ***operator elements*** are a finite set of actions that take a state of knowledge as input and produce a new state of knowledge as output. Users typically use a finite number of operators to change the knowledge states. PBG, which was grounded on the information processing theory [19], has been used in a large body of research to study the human problem solving process. Examples of the domains that have been studied include: financial analysis [5], mathematical programming [20], scheduling [14], and conceptual data modeling [3]. We found it useful to represent the online information retrieval process in terms of PBG.

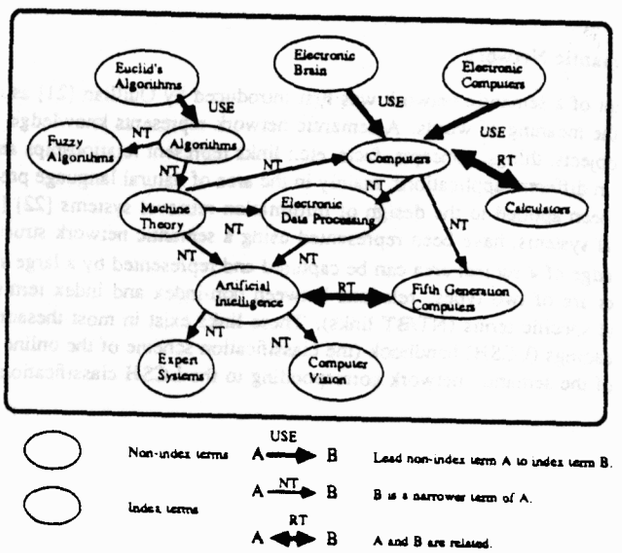


Figure 2: A sample network of LCSH terms

In the information retrieval context, the essential knowledge elements are the terms used to express the query. The operator elements are the moves or actions that change the content of the query. We refer to them as the *semantic operators*. As shown in Table 1, we observed five *semantic operators* used by searchers to move from one datum to the next. They are: ST (synonymous term operator, moves from a non-index term to its synonymous index term), BT (broader term operator, from specific to general term). NT (narrower term operator, from general to specific term), AT (adjacent term operator, from one term to another term which partially overlaps in meaning), and DT (disjointed term operator, from one term to another semantically different term). These operators were derived from our empirical studies. Details of how searchers used these operators in refining their queries are presented in Section 5.

In this subsection, we explain these operators in the context of the semantic network structure we proposed.

1. ST: The USE link in our semantic network leads from a non-index term to its synonymous index term. A searcher may follow this link to identify synonyms. These links are listed explicitly in the LCSH Handbook. For example, by following the USE link in Figure 2, a searcher can identify "Computers" as a synonym of "Electronic Brain."

2. BT or NT: The NT link leads from a broader term to a narrower term. BT link is the reverse of NT link. Following the NT (or BT) links we can determine the level of specificity of different terms. For example, by following the NT links in Figure 2, we know that "Electronic Data Processing" is more specific than "Computers," "Artificial Intelligence" is more specific than "Electronic Data Processing," and "Expert Systems" is more specific than "Artificial Intelligence." Both NT and BT links are transitive. Searchers may decide to broaden or narrow their queries by following these links.

3. AT: Terms which are not directly related via the USE or NT/BT links but have the same common ancestor (broader term) or descendant (narrower term) in the network are considered *adjacent* terms. We can determine the common ancestor or descendant of any two terms by activating all paths leading to/from the two terms in the network. The intersection of these paths is the common ancestor or descendant respectively. For example, "Fuzzy Algorithms" and "Machine Theory" have a common ancestor, "Algorithms," reached by following the reverse direction of NT links (BT links) in Figure 2 for both terms. "Electronic Data Processing" and "Machine Theory" are adjacent because they have a common descendant, "Artificial Intelligence." Searchers may change their topic of interest from an initial term to its "adjacent" terms.

4. DT: This operator represents the transition from one term to another term which is semantically disjointed (i.e., one term cannot reach another via the links in the network). This generally represents a change of query from one topic to another disjointed topic.

Based on these primitives, we can construct a PBG to describe the progression of the state of the query through the problem space. Each node of the graph represents a particular state of knowledge for the searcher, and each arc represents an operator that was applied to transit to the next state. By superimposing the PBG on the semantic network, we get an interesting visual picture of the trajectory of the query. Figures 4 and 5 demonstrate two query processes using this scheme. We discuss them in the next subsection.

### 4.2 Taylor's Theory for Online Query Refinement

In a seminal article on query refinement Taylor proposed four levels of query refinement that pertain to the client/information specialist interview session [27] [28]. These four levels are: visceral need, conscious need, formalized need, and compromised need. Changes of needs from one level to another are indicated in Figure 3. The librarian's role in assisting query refinement has been well recognized in the prior studies. We postulate

| Semantic Elements | Description |
|---|---|
| Knowledge Element: TERM | Terms used by the searcher. |
| Operator Element: | (Going from one term to another:) |
| ST | Synonymous Term |
| BT | Broader Term |
| NT | Narrower Term |
| AT | Adjacent Term |
| DT | Disjointed Term |

Table Semantic elements for query refinement

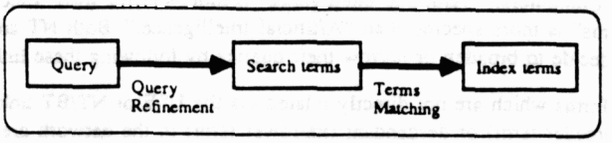


Figure 3: Taylor's theory of query refinement

that online infomation retrieval systems also assist searchers in refining or articulating queries. but possibly in a more passive way.

We represented the 30 searcher/system interactions in terms of the PBG and then analyzed these PBGs based on Taylor's model. Although our study was unable to identify the visceral need, it revealed the remaining three levels of needs. (However, the very existence of the searcher's query indicates the presence of the visceral need.) Definitions for the four types of needs in the online task environment an as follows [28]:

1. Q1. Visceral need: a vague son of dissatisfaction. or a certain incompleteness in the searcher'spicture of the world.
2. Q2. Conscious need: the conscious, within-brain description of the need. In our study, this was revealed in the searchers' think-aloud protocols during their online searches.
3. Q3. Formalized need: the formal statement of need. This is a searcher's query statement. In our study we elicited both pnscarch and postsearch query statements, using interviews.
4. Q4. Compromised need: the query as presented to the infomation system. This is disclosed by the actual search terms used by a searcher on the system. They were recorded in our log file.

The data collection techniques used in our study, namely, log file, tape-recording of the think-aloud protocols, and interviews, were useful for identifying the searchers' information needs at various levels. We observed numerous examples of query refinement from Q2 to Q3 and from Q3 to Q4.

Eight of the subjects changed the intent of their queries. Their postsearch statements of their information needs were significantly different from their presearch statements (a change in the formalized needs). These searchers' conscious needs emerged during the search and were formalized during the retrieval process. This query refinement process involved the formalization of the conscious needs (Q2-Q3).

Twenty-one (21) subjects' presearch and postsearch statements of information needs were unchanged (no changes in the formalized needs). However, the search terms used by these subjects varied during the search (although these terms were semantically close to the presearch and postsearch statements). This can be considered a process of forming a compromised need from a formalized need (Q3-Q4). The higher frequency of the second type of refinement (Q3-Q4) may have resulted from the searchers' attempts to provide search terms that the system could recognize and that represented their complete information needs. In the following subsections, we describe these two types of query refinement processes (formalize the conscious need, Q2-Q3, and compromise the formalized need, Q3-Q4) with examples.

### 4.2.1 Formalize the Conscious Needs: (Q2-Q3)

For eight of the subjects we observed, their presearch and postsearch query Statements were significantly different (in terms of contents). The subjects' think-aloud protocols revealed their conscious needs, which involved needs that were different from what had been stated in their prescarch statements. These conscious needs were formalized and then included in their postsearch statements.

Figure 4 illustrates how the content of one subject's query changed during the search process. The knowledge components (terms) involved in this interaction are represented using the semantic network structure we described earlier (see the ovals and the light arrows). The flow of the search process (in time sequence) is captured by a PBG (follow the heavy arrows from *Start* to End). The terms involved in the search are represented by the ovals in Figure 4. The concepts involved in the subject's presearch query statement (rectangular boxes) as well as the postsearch query Statement (shaded ovals) are also shown in Figure 4.

Six groups of terms were mentioned in this query. Each group consisted of terms that comprised a *concept*. We refer to this as a *concept group*. In terms of semantic network representation, words that address a similar concept are connected via some links (the ST or NT links as shown in Figure 4). In this example. six concept groups were involved: Group 1 - "error" and "error - psychology" (connected by an NT link as shown in the upper-right hand corner of Figure 4). Group 2 - "routine work" and "motor skills" (connected by an NT link). Group 3 - "data," "data entry," "data editing." "data processing," "data processing service center." "data processing service center - management." and "data reduction" (connected by the NT links). Group 4 - "ergonomics" and "human factors" (connected by an ST link). Group 5 - "learning," and Group 6 - "performance." The subject's presearch query statement was very fuzzy, "how people make errors." It mentioned only one concept in Group 1 (indicated by the rectangular box in Figure 4). Five new concepts were introduced during the search (Groups 2, 3.4, 5, and 6). After the search, the searcher's query statement was: "the human factors in the routine work of data entry." It contained concepts in Groups 2, 3. and 4 (see the shaded ovals in Figure 4). We postulate that the searcher's own brainstorming and the incidental information displayed by the system resulted in a query refinement process similar to the one observed in the client/librarian interview session. However, because the searchers had little support from the system. this query refinement process was often time-consuming and ineffective.

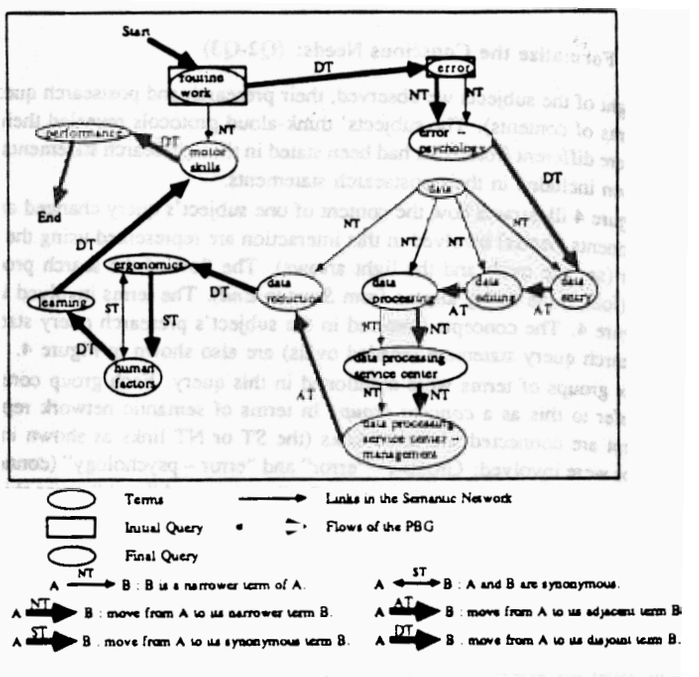


Figure 4: An example of formalizing the conscious needs

### 4.2.2 Compromise

The contents of 21
query statements. H
information system
the result of the sear

In terms of sem
(ST) in order to gen
narrower terms (NT
partially in meaning

Figure 5 shows
terms of its meaning
a number of terms
and postsearch state
the rectangular box
the searcher tried o
"multi-level plannir
to **End** in Figure 5

There may be s
document indexing
forced to use diffe

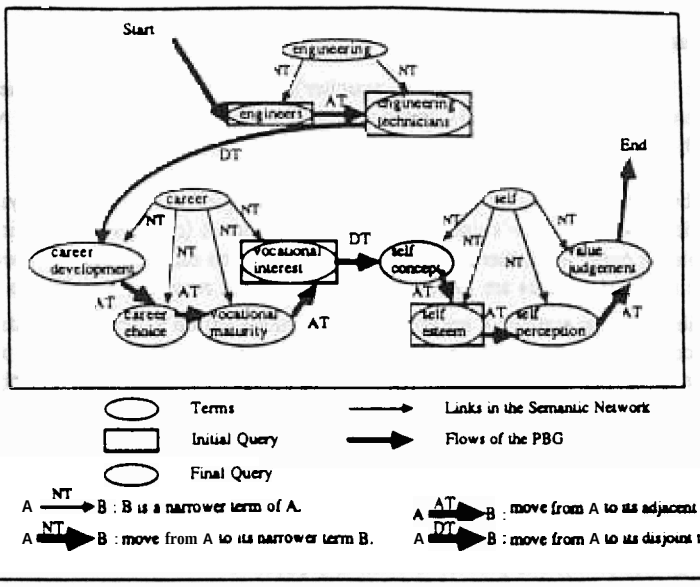


Figure 5: An example of compromising the formalized needs

### 4.2.2 Compromise the Formalized Needs: (Q3-Q4)

The contents of 21 subjects' presearch query statements had the same or similar meaning as their postsearch query statements. However, the search **urms** they used during the search vaned. The search terms input to the information system (the compromised needs) were variations of their formalized needs. We posit that this was the result of the searchers' attempts to adjust to the system's vocabulary.

In terms of semantic network representation. searchers may move from one term to its synonymous terms (ST) in order to generate matches; to its broader terms (BT) in order to broaden the scope of their queries; to its narrower terms (NT) in order to focus their searches; or switch from one topic to another topic which overlaps partially in meaning (AT) in order to cover different aspects of their queries.

Figure 5 shows an example of how one searcher chose different search terms to express an unchanged (in terms of its meaning) query. Only one concept group was involved in this search, but this concept group contains a number of terms which are related to "planning" (see the ovals in Figure 5). The subject's prescarch statement and postsearch statement addressed the same information need. "hierarchical planning" (see the shaded oval and the rectangular box in Figure 5). although different terms were chosen to express the query. After fading initially. the searcher tried other semantically close terms to generate matches. These terms included: "business planning." "multi-level planning." "large scale planning." and "organizational planning." (Follow the dark arrows from Start io End in Figure 5.)

There may be several hypotheses to explain this process. First, because of the controlled vocabulary used for document indexing, search terms used initially may not have produced any matches. So searchers were often forced to use different terms. Second, information displayed by the system may have suggested new clues for

the search. Lastly. searchers may have found it more prudent to approach their queries from different angles (by using synonymous, broader. narrower, or adjacent terms).

### 4.3 Query Refinement: Librarian vs Retrieval System

In earlier studies. the librarian's function in refining searcher's queries has been well documented. In this section we discuss the differences between the way librarians assist in refining queries and the way online retrieval systems provide such help.

1. The Role: Librarians assume a more active role in refining queries than retrieval systems do. Librarians attempt to understand the patron's underlying information needs (conscious needs) in a very early stage. This active role in query refinement is useful for achieving an efficient (minimal wasted effort) search. Current online retrieval systems are too passive in this respect, requiring searchers to do too much.

   Experienced librarians, in particular, are good at identifying a patron's conscious needs by detecting inconsistency or incompleteness in the patron's queries. They play an important role in prompting patrons to formalize their conscious needs. searchers of online systems. on the other hand, rarely have this support.

2. Sources of Knowledge: In searcher/librarian interactions. both the searcher's subject area knowledge and the librarian's subject area, classification scheme. and system knowledge contribute to the generation of new cues for the search. During searcher/system interactions. on the other hand, searchers *can* rely only on their own knowledge in the subject area (searchers in general have limited classification scheme and system knowledge) and incidental information displayed on the screen. For searchers who have little subject area knowledge (e.g., freshmen), this process is often unproductive.

   Librarians' classification scheme knowledgc is often used to assist patrons in generating queries the information systems can recognize (compromised needs). Their knowledge about the system's controlled vocabulary, the cross-referencing structure, and the index principles employed helps patrons obtain a good surrogate for their formalized needs. In the searcher/system interactions. this process is essentially one of trial and error.

3. User Modeling: One capability which we found to exist only in the librarians was their ability to perform a user modeling function. For example, librarians may expect vagueness in the queries from inexperienced and less sophisticated searchers (e.g., freshmen). Librarians generally spend a considerable amount of time in sharpening queries from these clients. But when dealing with more sophisticated searchers (e.g., Ph.D students), they may assume the searchers know exactly what they what. This user *modeling* function of the reference librarians has been reported in [7].

We summarize these differences between librarian and online system assistance in refining queries in Table 2. It is clear that in order to design more effective and useful online reuieval systems, query refinement functionality needs to be incorporated into the retrieval systems.

## 5 Process Model of Online Query Refinement

We derived some interesting empirical findings penaining to how searchers refine their queries when using the reuieval system. These findings were used io develop a process model that can facilitate online query refinement.

### 5.1 Two Approaches to Or

We observed that searchers refu
browsing and retrieval by inst

#### 5.1.1 Semantic-Based Browsi

Searchers may obtain their *sea*
studies we identified a typology
included synonymous term (ST)
(DT). The searcher's query refu
of concepts (terms. topics, etc.)
We discuss these semantic oper

1. Synonymous Term (ST):
   terms were self-generated
   as a consequence of havi
   term. "ergonomics," to ex

2. Broader Term (BT): Sea
   *are* more general. This c
   under the specific terms s
   looking for books about '
   to "statistics" after obtain

3. Narrower Term (NT) So
   terms. This occurred w
   expectation of what a re:
   huge set of matched citat
   "measurement." Searcher

4. Adjacent Term (AT): Sc
   initial terms in content.
   These adjacent terms ma

| Differences | Searcher/Librarian | Searcher/System |
|---|---|---|
| The Role | o Active | o Passive |
| Sources of Knowledge | o Subjca Am. System, Classification Scheme | o Subject Area |
| User Modeling | o Yes | o No |

Table 2: Librarian vs. system in query refinement

### 5.1 Two Approaches to Online Query Refinement

We observed that searchers refined their queries using two approaches to which we referred as semantic-based browsing and retrieval by instantiation.

#### 5.1.1 Semantic-Based Browsing

Searchers may obtain their search terms by browsing the semannc network of concepts. From our empirical studies we identified a typology of *semantic operators* that searchers used for refining heir queries. This typology included synonymous term (ST), broader term (BT). narrower term (NT), adjacent term (AT), and disjointed term (DT). The searcher's query refinement process can be viewed as a traversal (browsing) in this semantic network of concepts (terms, topics. etc.) using these five operators. Figures 4 and 5 illustrate this type of query refinement. We discuss these semantic operators below.

1. Synonymous Term (ST): In the searcher/system interactions, we observed a proccss where the synonymous terms were self-generated by the searchers when their initial terms matched too few citations. For example. as a consequence of having found no matches by using "human factors." a user next used a synonymous term. "ergonomics." to express his query.

2. Broader Term (BT): Searchers may change from terms which an more specific in meaning to terms whch are more general. This change may be due in part to a searcher's misconception that citations classified under the specific terms should also be classified under broader terms. For example. a searcher who was looking for books about "statistical power" immediately changed her search term from "statistical power" to "statistics" afwr obtaining no matches from the first term.

3. Narrower Term (NT) Sometimes. searchers may decide to narrow their queries by choosing more specific terms. This occurred when a set of matched citations was too large. Searchers often had their own expectation of what a reasonable number of matched citations should be. For example. after deriving a huge set of matched citations under "statistics," a searcher immediately returned to a more specific term. "measurement." Searchers also used narrower terms when the initial term generated irrelevant citations.

4. Adjacent Term (AT): Searchers may change their search topic to one which partially overlaps with their initial terms in content. This occurred when few relevant citations were derived from their initial terms. These adjacent terms may capture aspects of the query which were not represented in the initial terms.

For example. a searcher switched from "economics of informdon" to "game theory" in two consecutive search stages in an attempt to explore the different aspects of the "bargaining problem."

5. Disjointed Term (DT): Searchers may change their search topic to another semantically disjointed topic. This occurred when multiple topics were involved in a query and when previously ignored conscious needs were revealed during the search. In an online retrieval system that has full Boolean capability. this type of query can be expressed using Boolean operators. Searchers. however. may have problems using these operators. In our study. four subjects attempted to use a Boolean search. while most subjects searched with single term (not using the Boolean search option) even though they had more than one topic in their query. For instance, a searcher used "hypertension" and "salt substitute" (two disjointed topics) in two separate steps in an effort to search for materials about "the effect of salt substitutes on hypertension" This query can be expressed by the Boolean logic: "hypertension AND salt substitute."

These five operators correspond to the various links of the semantic network structure we have proposed. The ST operators follow the USE links in the semantic network (see Figure 2). The NT and BT operators follow, in either direction, the NT or BT links. Terms which have a common ancestor or descendant (but not on a hierarchy of NT or BT links) are considered *to* have an AT relationship (going from one such term to another is considered as an AT operation). Terms which are not linked in the network are considered to have an DT relationship (likewise. going from one such term to another is considered as an DT operation). We believe that an online thesaurus that is based on this semantic network structure can help searchers refine their queries.

### 5.12 Retrieval by Instantiation

Detailed citation information can also help searchers obtain new cues for search. This includes: the title, the author, the publisher, and the index term of a book. In our studies. we frequently heard statements such as:

This book is exactly what I am looking for

I am looking for books similar *to* this one.

By using information associated with the citations that are right on target, searchers can obtain other relevant citations. Searchers can use the index terms derived from the matched citations *to* perform a subject search. New relevant citations can be obtained. Searchers can examine the titles of the matched citations in order to elicit new search terms (tide also reflects the content of a book). Searchers can find all books written by a particular author in the area (most authors work in a few specific areas). Searchers can sometimes find all new books published by a particular publisher (many publishers specialize in certain subject mas). This retrieval mechanism. which we referred to as retrieval by instantiation. has also been found useful in the design of other informadon retrieval systems [29] [30] [32].

## 5.2 A Process Model

Grounded on our empirical findings, we have developed a proccss model for assisting query refinement using online information retrieval systems. This process model consists of five stages as shown in Figure 6. They arc: Initial Query Stage, Terms Grouping Stage. Relevance Evaluation Stage, Terms Solicitation Stage, and Citations Instantiation Stage. We discuss each of these stages below.

### 5.2.1 Initial Query Stage

A search session starts with the Initial Query Stage (see the box at the top of Figure 6). Two activities are involved in this stage. First. searchers express their queries by using a few search terms. These search terms represent the searchers' informatic terms.

Second, the search terms elicit the LCSH Handbook). Non-index the USE links in this thesaurus. T operators) whch represent the sea

### 5.2.2 Terms Grouping Stage

The Terms Grouping Stage CON derived from the previous stage acc network) and generating new canc rank these newly derived terms.

1. Grouping: Grouping of tem lead to/from these terms. Th via some paths arc classifiec relating to a dissertation top

   An important by-product oft paths. These new terms can

2. Ranking: The concept grou terms (originators) in each g than a group with fewer on

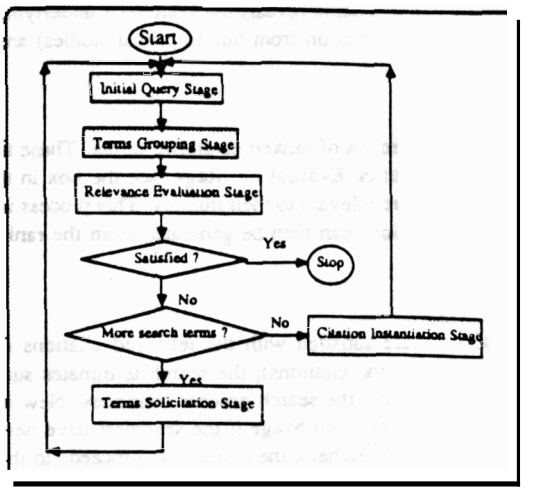


Figure 6: A proccss model for query refinement

represent the searchers' information needs. The searchers can use Boolean operators to combine their search terms.

Second, the search terms elicited are used *to* consult a semantic network-based online thesaurus (based on the LCSH Handbook). Non-index terms supplied by the searchers can be translated to index terms by tracing the USE links in this thesaurus. The output of this stage is a list of index terms (possibly combined by Boolean operators) which represent the searchers' qucrics.

### 5.2.2 Terms Grouping Stage

The Terms Grouping Stage consists of two processes. The first proccss focuses on grouping the index terms derived from the previous *stage* according to their contents (relationships *to* each other with respect to the semantic network) and generating new candidate terms from each group. The second process applies a few heuristics *to* rank these newly derived terms.

1. Grouping: Grouping of terms is accomplished by first instantiating all paths in the semantic network which lead to/from these terms. This operation is referred to as *spreading activation*. Terms which *arc* connected via some paths are classified into the same concept group. Complex qucrics (e.g., a Ph.D. student's query relating to a dissertation topic) may often involve more than one concept group.

   *An* important by-product of this spreading activation proccss is a set of new terms found along the activated paths. These new terms can become good candidate terms for the searchers' qucncs.

2. **Ranking: The** concept groups derived arc first ranked according to the total number of searcher-supplied terms (originators) in each group. A group with many originators is considered more important to the query than a group with fewer originators. We refer to this heuristic of ranking as the *principle of origination*.

Next, we rank the newly derived terms within each group. The ranking is based on the *specificity* principle of A term which is more specific reveals the searcher's underlying information need better than its more general counterparts (an observation from our empirical studies) and thus should be ranked higher.

#### 5.2.3 Relevance Evaluation Stage

The Terms Grouping Stage produces groups of ranked candidate terms. These terms need to be evaluated by the searchers, however. During the Relevance Evaluation Stage (see the box in the center of Figure 6), searchers select and rank those terms they think are relevant to their queries. This process is intended to reveal the searcher's conscious needs. A list of ranked citations can then be generated from the ranked candidate terms.

#### 5.2.4 Terms Solicitation Stage

As shown in Figure 6, if searchers are satisfied with the retrieved citations (both in terms of the number of matched citations and the relevance of the citations), the search terminates successfully. On the other hand, if searchers are not satisfied with the results, the search process continues. New information needs to be supplied. The search moves on to the Terms Solicitation Stage if the searchers have new search terms to provide. When no search terms can be supplied by the searchers, the interaction proceeds to the Citations Instantiation Stage. This stage involves the use of the detailed information in the retrieved citations.

At the Terms Solicitation Stage searchers supply new search terms. These new terms are first translated into index terms. A *spreading activation* process using these new terms and the previous terms then follows. New candidate terms and citations can be generated. This iterative process ends when no new terms can be supplied by the searchers. This stage attempts to simulate the searchers' semantic-based browsing approach described earlier.

#### 5.2.5 Citations Instantiation Stage

While the Terms Solicitation Stage aims at soliciting new terms from the searchers, the Citations Instantiation Stage instantiates the information embedded in the citations that have been selected by the searchers. It attempts to simulate the retrieval by instantiation approach for query refinement. The index terms assigned to the selected citations, in particular, provide good clues for the search. By performing a *spreading* activation process using the old terms and the index terms just derived, new relevant terms and citations can be obtained.

As shown in Figure 6, our proposed process model for online query refinement is iterative in nature. By applying the heuristics and the refining mechanisms in this model repeatedly, a retrieval system can better assist searchers in articulating and refining their queries.

## 6 Conclusion

Electronic information storage and retrieval systems have changed the way people retrieve information. Investors access financial data of companies via their terminal at home. Lawyers consult law cases by online browsing of databases. Researchers obtain information about relevant studies by using online catalogs and online bibliographic databases. While the amount of information available for online access increases dramatically, searches are often problematic. This may be due in large part to the difficulty involved in articulating and refining the searchers' underlying information needs. Human information specialists' role in refining queries has been well recognized in prior studies. In our research, we observed an online query refinement process consistent with Taylor's theory.

However, the process was passive ar
were present.

We observed that searchers refine
and at the citation level using author
representation to capture the subject
model, we believe, can serve as a ba
We have incorporated this model in
referred to [8].

## References


[1] W. Ross Ashby. *An Introducti*

[2] Marcia J. Bates. Subject acc
*Information Science*, 37(6):35

[3] D. Batra and J. G. Davis. A stu
between expert and novice de
*Systems* (/CIS-89). pages 91–1

[4] N. J. Belkin, R. N. Oddy, and
*Journal of Documentation*, 38

[5] J. Marinus Bouwman. Huma
*Management Science*. 29:653-

[6] R.J. Brachman. what's in a c
*Man-Machine Studies*. 9, 197

[7] Hsinchun Chen and Vasant Dh
interaction In *Proceedings* o

[8] Hsinchun Chen and Vasant I
systems. In *Proceedings of t*

[9] Paul R. Cohen and Rick Kj
networks. Information *Proce*

[10] Timothy C. Craven. Thesa
permuted index displays. *Inf*

[11] K. Anders Ericsson and He
Cambridge, Massachusetts,

[12] G. W. Furnas. T. K. Landau
communication. *Communica*

[13] M. D. Good. J. A. Whitesid
*cations of the ACM*, 27(10):

hiring is based on the *principle of* ying information need better than its and thus should be ranked higher.

ie ums need to be evaluated by the n the center of Figure 6). searchers ss is intended to reveal the searcher's anked candidate terms.

is (both in terms of the number of successfully. On the other hand. if w infomation needs to be supplied. new search terms to provide. When the Citations Instantiation *Stage.* ons.

e new terms are first translated into previous terms then follows. New when no new terms can be supplied based browsing approach described

archers, the Citations Instantiation elected by the searchers. It attempts index terms assigned to the selected *spreading activation* process using ions can be obtained.

finement is iterative in nature. By , a retrieval system can better assist

eople retrieve information. Investors sult law cases by online browsing of ine catalogs and online bibliographic ases dramatically, searches are often iculating and refining the searches' ng queries has been well recognized cess consistent with Taylor's theory.

lists

ors ork his ns. are

representation to capture the subject area knowledge and developed a process model for query refinement. This model, we believe, can serve as a basis for designing more "intelligent" and useful information retrieval systems. We have incorporated this model into the design of a knowledge-based document retrieval system. Readers


[2] Marcia J. Bates. Subject access in online catalog: a design model. *Journal of the American Society of Information Science,* 37(6):357–376, November 1986.

[3] D. Batra and J. G. Davis. A study of conceptual data modeling in database design: similarities and differences between expert and novice designers. In *Proceedings of the 10th International Conference on Information* System *(ICIS-89),* pages 91–100, Boston. MA, December 1989.

[4] N. J. Belkin, R. N. Oddy, and H. M. Brooks. Ask for information retrieval: Part I, background and theory. *Journal of Documentation,* 38(2):61–71, Junc 1982.

[5] J. Marinus Bouwman. Human diagnostic reasoning by computers: an illustration from financial analysis. *Management Science.* 29:653–672, Junc 1983.

[6] R.J. Brachman. what's in a concept: Structural foundations for semantic network. *International Journal of Man-Machine Studies,* 9. 1977.

[7] Hsinchun Chcn and Vasant Dhar. Reducing indeterminism in consultation: a cognitive model of user/librarian interaction In *Proceedings of the 6th National Conference on Artificial Intelligence (AAAI-87),* 1987.

[8] Hsinchun Chen and Vasant Dhar. A knowledge-based approach to the design of document-based retrieval systems. In *Proceedings of the 5th Conference on office Information Sysrem,* Cambridge, MA 1990.

[9] Paul R. Cohen and Rick Kjeldsen. Information retrieval by constrained spreading activation in semantic networks. *Information Processing and Management,* 23(4):255–268, 1987.

[10] Timothy C. Craven. Thesaural relations in a concept-network management system for customizing of permuted index displays. *Information Processing and Management,* 20(5/6):633–610, 1984.

[11] K. Anders Ericsson and Herbert A. Simon. *Protocol analysis: verbal reporr as &a.* The MIT Press, Cambridge, Massachusetts. 1984.

[12] G. W. Furnas. T. K. Landauer. L. M. Comer, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of rhe ACM,* 30(11):964–971, November 1987.

[13] M. D. Good, J. A. Whiteside, D. R. Wixon, and S. J. Jones. Building a user-derived interface. *Communications of rhe ACM,* 27(10):1032–1043, October 1984.

[14] B. Huguenard, M. J. metula, and F. J. Lerch. Performance ≠ behavior: a study in the fragility of expertise. *Proceedings of the 10th International Conference on Information Systems (ICIS-89)*. pages 101–117, Boston MA, December 4-6 1989.

[15] J. Jacoby and V. Slamecka. *Indexer Consistency Under Minimal Conditions*. Documentation, Inc., Bethesda MD, 1962.

[16] F. W. Lancaster. *Information Retrieval Systems*. John Wiley and Sons. Inc., 1979.

[17] Karen Markey. Levels of question formulation in negotiation of information need during the online presearch interview: a proposed model. *Information Processing and Management*, 17(5):215–225, 1981.

[18] Karen Markey and Pauline Atherton. *Online Training and Pracricc Manual for ERIC Data Base Searchers* Syracuse, *NY:* ERIC Clearinghouse on Information Resources, W C : ED 160109, 1978.

[19] A. Ncwcll and H. A. Simon. *Human Problem Solving*. Prentice-Hall, Englewood Cliffs. NJ. 1972.

[20] Wanda Orlikowski and Vasant Dhar. Imposing structure on linear programming: an empirical analysis of expert and novice models. In *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI-86)*, Philadelphia, PA, August 1986.

[21] M. R. Quillian. Semantic memory. In *Semantic Information Processing*, 1968.

[22] Peretz Shoval. Principles, procedures and rules in an expert system for information retrieval. *Information Processing and Management*, 21(6):475–487, *1985.*

[23] R.F. Simmons. *Semantic Network: Their Computation and Use for Understanding English Sentences*. In Schank, R.C. and Colby, KM. (eds.), Computer Models of Thought and Language, Freeman. 1973.

[24] Mary Elizabeth Stevens. *Automatic Indexing: A State-of-the-art Report*. U.S. Government Printing Office Washington, DC. 1965.

[25] Gary W. Strong and M. Carl Drott. A thesaurus for end-user indexing and retrieval. *Information Processing and Management*, 22(6):487–492, 1986.

[26] R. Tagliacozzo and M. Kochen. Information-seeking behavior of catalog users. *Information Storage and Retrieval.* 6:363–381, 1970.

[27] Rober S. Taylor. The process of asking questions. *Am. Documen.*, 13:391–396, 1962.

[28] Rober S. Taylor. Qucsuon-negotiation and information seeking in libraries. *College and Research Libraries*, 29:178–194, May 1968.

[29] Bemd Teufel. Natural language documents-indexing and retrieval in an information system. In *ICIS'88 Conference Proceedings,* Minneapolis. MN, December 1988.

[30] F. N. Tou. M. D. Williams. R. Fikes, A. Henderson. and T. Malone. Rabbit: *An* intelligent database assistant. In *Proceedings of the National Conference on Artificial Intelligence*. 1982.

[31] D.A. Waterman and A. Newell. Protocol analysis as a task for artificial intelligence. *Artificial Intelligence* 2:285–318, 1971.

[32] Michael David Williams 352, 1984.

[33] W.A. Woods. *What's in* Representation and Und

study in the fragility of expertise. In
*(ICIS-89)*.pages 101–117, Boston

*ons*. Documentation. Inc., Bethesda,

Inc., 1979.

ion need during the online presearch
, 17(5):215–225, 1981.

*nual for ERIC Data Base Searchers*.
3D 160109, 1978.

Englewood Cliffs, NJ. 1972.

rogramming: an empirical analysis
*Conference on Artificial Intelligence*

, 1968.

r information reuicval. *Information*

*nderstanding English Sentences*. In
id Language, Freeman. 1973.

*t*. U.S. Government Printing Office.

and retrieval. *Information Processing*

dog users. *Information* Storage *and*

391–396, 1962.

ries. *College and Research Libraries*.

an information system. In *ICIS'88*

ne. Rabbit: *An* intelligent databaw
*ligence*. 1982.

al intelligence. *Artificial Intelligence*.

[32] Michael David Williams. What mikes rabbit run? *International Journal of Man-Mochinc Studies*. 21:333–352. 1984.

[33] W.A. Woods. *What's in a Link: Foundations for Semantic Networks*. In Bobror, G. and Collins. A. (eds.), Representation and Understanding: Studies in Cognitive Science, Academic Press, New York. NY, 1975.