User Misconceptions of Information Retrieval Systems

User Misconceptions of Information Retrieval Systems

Hsinchun Chen
Department of Management Information Systems, University of Arizona
Tucson, Arizona 85721, USA
Vasant Dhar
Department of Information Systems, New York University
New York, New York 10003, USA

Summary

We report results of an investigation where thirty subjects were observed performing subject-based search in an online catalog system. The observations have revealed a range of misconceptions users have when performing subject-based search. We have developed a taxonomy that characterizes these misconceptions and a knowledge representation which explains these misconceptions. Directions for improving search performance are also suggested.

Table of Contents

1. Introduction

While archival information sources such as libraries are becoming more computerized, access to such information is often difficult because of the indeterminism involved in the process by which documents are indexed, and the latitude users have in choosing terms to express a query.

Most online catalog systems offer capabilities for access using what are referred to as *known items* such as author, title, and call number, and *non known items* such as subject area. While known-item search is easy to support, subject-based search can be difficult, often requiring the assistance of a reference librarian. For these types of queries, the problem of finding relevant documents 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 pertinent to the information storage and retrieval system.

Our overall research goals are twofold. The first goal is to understand how reference librarians help users with subject search. The second goal is to understand the problems users have in trying to do subject-based search. In a previous paper (Chen & Dhar, 1987), we described a model of librarian-user interaction that showed the strategies employed by librarians to structure subject-based queries. In this paper, we report results of an investigation where thirty subjects were observed performing this type of search. The observations have revealed a range of misconceptions users have when performing this type of search. The specific results we report in this paper are:

1. a taxonomy that characterizes these misconceptions,
2. hypotheses of the causes of misconceptions,
3. a knowledge representation scheme that explains these misconceptions, and
4. directions for alleviating the problems we have observed in subject-based search.

In addition, we have developed an integrated framework for classifying previous research and situating our work within this framework.

2. A Framework for Information Retrieval

The importance and difficulties of subject-based search have been well documented. In one study of card catalog system, subject-based search was found to constitute about 40 percent of catalog use (Markey, 1984). Another study of online catalogs showed that subject-based search constituted between one-third and a one-half of all searches (Rochell, 1984). Even though keyword matching capabilities have been incorporated into many online catalog systems, the need for additional subject-related search capabilities has headed the list of desired improvements in several studies (Matthews, Lawrence & Ferguson, 1983; Kaske & Sanders, 1980; Larson & Graham, 1983).

In this section, we present a theoretical framework for understanding the information storage and retrieval process in terms of the "agents" involved in the indexing and searching processes. This framework as shown in Figure 1 shows the human agents involved in these processes, the types of knowledge these agents possess, and the observed characteristics of their indexing and searching behaviors. This framework is described in the remainder of this section.

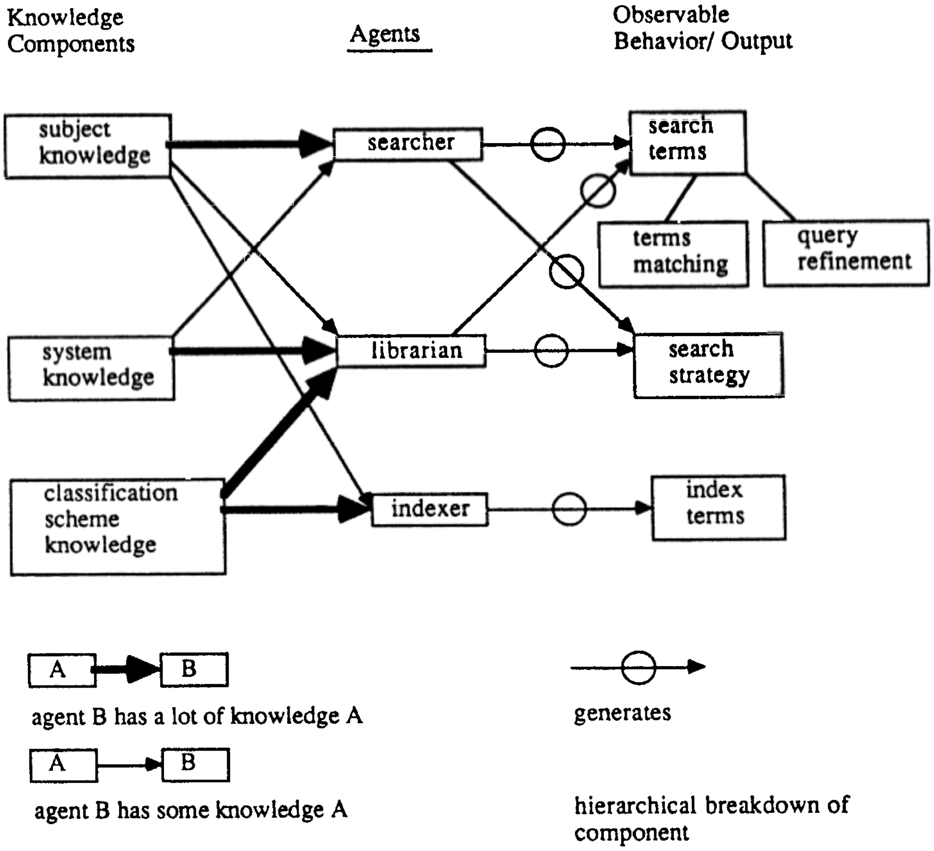


FIG. 1. A framework for information storage and retrieval

2.1. Knowledge Components

As noted at the outset, three types of knowledge are involved in subject-based search. First, classification scheme knowledge that is used for indexing documents is also required in searching for them. Secondly, subject area knowledge is required for expressing a query in appropriate terms. Lastly, system knowledge is necessary. These three knowledge components are presented towards the left of Figure 1. These components will be discussed in detail in Section 4.

2.2. Agents

The three types of knowledge outlined above are typically distributed among three parties which we refer to as "human agents". These agents include indexers, who classify the documents based on some predetermined classification scheme; searchers, who express their queries using some terms; and reference librarians, who serve as the intermediaries between searchers and retrieval systems.

Searchers generally do not have classification scheme knowledge. Their knowledge of the subject area and the system functionality varies widely. Indexers generally have a lot of classification scheme knowledge and some subject knowledge (they are, however, not involved in the retrieval process). 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 in Figure 1 using links characterizing "strong" and "weak" knowledge.

2.3. Observable Behavior/Output

Indexing uncertainty and search uncertainty are the primary sources of information retrieval problems. Indexing uncertainty arises because different expert indexers can assign different index terms for a given document (see the box labelled "index terms" in Figure 1). Search uncertainty arises because searchers have latitude in choosing terms to express a query (see the box labelled "search terms" in Figure 1) and the search strategies they employ in acquiring information (see the box labelled "search strategy" in Figure 1).

Because of the indeterminism involved in indexing and searching, an exact match between the searcher's terms and those of the indexer is unlikely. This is referred to as the terms matching problem (see the box labelled "terms matching" in Figure 1). Secondly, the search terms used may not in fact represent what the searcher is really looking for (assuming for the moment that he knows what he wants). This is referred to as the problem of query refinement (see the box labelled "query refinement" in Figure 1). In the remainder of this section, we discuss each problem in detail in the context of prior research.

2.3.1. Indexing Uncertainty

The process of indexing is partly indeterminate. Evidence suggests that different indexers, well trained in an indexing scheme might assign index terms for a given document differently. It has also been observed that an indexer might use different terms for the same document at different times (Jacoby & Slamecka, 1962; Stevens, 1965).

Several approaches have been proposed by researchers aiming at improving the indexing of documents. These include: use of the Dewey Decimal Classification (Cochrane & Markey, 1985), more extensive linkages between fields in different records that allow users to browse and navigate through a database (Noerr & Bivins Noerr, 1985), and the application of the "hypertext" concept to catalogs, that is, breaking the linearity of the traditional file structure and providing links in a variety of different directions in records (Hjerppe, 1985).

In our research, we take indexing uncertainty as given. We focus only on improving the search process, assuming that some level of indexing uncertainty will always exist.

2.3.2. Search uncertainty

Search uncertainty refers to the latitude searchers have in choosing search terms.

A. Search Term

1. **Terms Matching:** A high degree of uncertainty with regard to search terms has been observed. Searchers 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 less than 20 percent (Gomez & Lochbaum, 1984; Good, Whiteside, Wixon & Jones, 1984; Furnas, Landauer, Gomez & Dumais, 1987). This limits the success of various design methodologies for controlled vocabulary-driven interaction (Furnas et al., 1987).

   Bates (1986) argues that for a successful match, the searcher must somehow generate as much "variety" (in the cybernetic sense, as defined by Ashby (1973)) in the search as is produced by the indexers in their indexing. The variety produced by an indexer can also be viewed as redundancy in the sense that it consists of partially overlapping classifications applied to a document. To increase the chances of a successful match, there should be a number of indexes for each document. This requires preserving the redundancy (generated by the indexer) associated with each document. In practice, however, catalog systems discourage redundancy for the following reasons (Bates, 1986; Chan, 1986):

   a) Whole document indexing: The cataloger working according to the Library of Congress or some other scheme is trained to index the whole document, not parts or concepts within it.

   b) Uniform Heading: The principle of uniform heading holds that for any description there is to be one and only one heading reflecting that description.

   c) Specific Entry: Each document is to be entered under a category (heading) which is specific to the content of the book, neither broader or narrower in scope than the scope of the book's contents.

d) Limited cross-reference structure: cross references are frequently an afterthought to "augment" the basic catalog organization (Bates, 1977).

In summary, these tendencies to reduce redundant access points, decreasing the likelihood that a user will generate the right term for retrieval.

2. **Query Refinement:** While indexers use the rule of specificity for indexing, users tend to approach a search by specifying broader terms first. There might be several reasons for this. One hypothesis is that users often do not have "queries", but what Belkin calls an "anomalous state of knowledge" (Belkin, Oddy & Brooks, 1982). Users often expect to refine this anomalous state into a query, *through* an interactive process. The organization of a catalog or a system, however, does not always facilitate this type of query refinement. In contrast, reference librarians appear to be particularly adept at this function. Taylor suggests that user'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). The actual query presented to the information system, however, may be compromised due to the user's expectation of the system (compromised need) (Taylor, 1968). Based on Taylor's model, a similar model for describing query refinement during the pre-search interview between the reference librarians and the online searchers was developed by Markey (1981). The importance of query refinement during the information retrieval process is well recognized in the prior research.

   Attempts on helping users refine their queries in the online retrieval systems have been made. An online thesaurus can help articulate the user's query by presenting the terms neighboring to the search terms semantically and asking the user to select the appropriate terms (Shoval, 1985). This approach addresses the query refinement problem at the subject heading level. The other approach for query refinement is "retrieval by instantiation". To pose a query, the user interactively constructs a description of his target items by evaluating successive examples and counterexample instances (database records). Information which is irrelevant to the users' query is removed, whereas information which is missing in the instances will be included. This approach is presented in a database project called RABBIT (Tou, Williams, Fikes, Henderson & Malone, 1982; Williams, 1984).

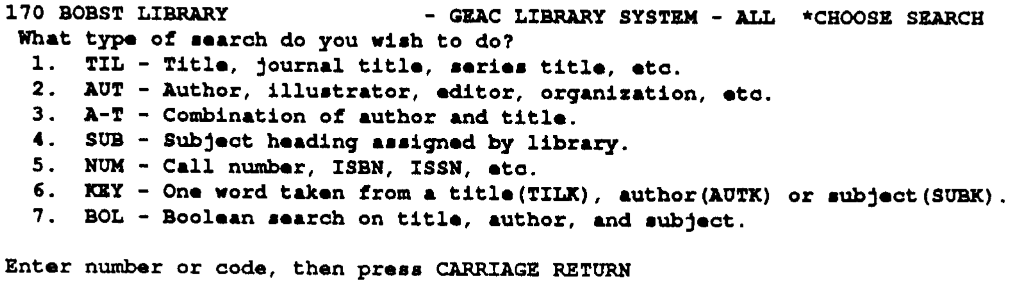
B. Search Strategy

Despite the fact there is considerable latitude involved in the term a searcher may employ to describe a subject area, the approach adopted by the users for performing search varies. Search strategy is often used to describe the plan or approach for the whole search, whereas search tactic refers to a move or maneuver made to further a search (Bates, 1979). Bates has described 29 tactics that are used in information searching. These tactics are grouped into four categories: 1) monitoring tactics are actions to keep the search on track, 2) file structure tactics are techniques for traversing the information within information system, 3) search formulation tactics are the tactics to aid in the process of designing or redesigning the search formulation, and 4) terms tactics are the moves to select and revise search terms. These tactics are applicable in both manual and online systems. In a card catalog study, two strategies for searching have been identified: a "self-reliant" style where users generate their own search terms and a "catalog-oriented" style where users use the terms found in the card catalog (Tagliacozzo & Kochen, 1970). 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 (Lancaster, 1979).

In the remainder of this paper, we present results of our study using the general framework developed in this section. Specifically, we focus on the searcher, identifying and categorizing the various types of misconceptions we have observed in the retrieval process.

3. Research Design

The NYU online catalog system, Bobcat, lists over 600,000 catalog records including all new materials purchased after 1973 and many older items previously listed in the card catalog; journals are not listed under Bobcat. The system provides seven search options, namely, title search, author search, combination of author and title search, subject search, number search, keyword search, and Boolean search. These options are available in most online catalog systems. Figure 2 shows the initial screen of the Bobcat system.



```
170 BOBST LIBRARY                  - GEAC LIBRARY SYSTEM - ALL  *CHOOSE SEARCH
 What type of search do you wish to do?
  1.   TIL - Title, journal title, series title, etc.
  2.   AUT - Author, illustrator, editor, organization, etc.
  3.   A-T - Combination of author and title.
  4.   SUB - Subject heading assigned by library.
  5.   NUM - Call number, ISBN, ISSN, etc.
  6.   KEY - One word taken from a title(TILK), author(AUTK) or subject(SUBK).
  7.   BOL - Boolean search on title, author, and subject.

Enter number or code, then press CARRIAGE RETURN
```

FIG. 2. The search options in Bobcat

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 interest. In general, the most frequently chosen option was 4 above, followed by 6. But there were also subjects using known-item search options to perform a subject-based search. That is, they used the title (or portion of the title), author, or call number to find documents within certain subject areas.

At the beginning of the interaction subjects were asked to write down briefly, what they were looking for. Subjects were also asked to think aloud during the interaction. This protocol was tape-recorded, and the interaction between the user and the system was logged. Interactions lasted between 5 and 40 minutes, with a median of 15 minutes. After the interaction, subjects were asked a few follow-up questions pertaining to the search process and the problems encountered during search. Suggestions were also elicited about how the system might be improved.

4. A Taxonomy of Misconceptions

We define a misconception as one where a user performs an erroneous action, uses erroneous terms, or goes about achieving a goal using an erroneous or suboptimal procedure. Under this definition, a lack of knowledge about something (ignorance) is also treated as misconception. We acknowledge that this definition stretches slightly the true meaning of misconception, but adopt it anyway for lack of a better term.

The logs (and to a small extent the protocols) revealed between one and seven misconceptions per user. Some of these misconceptions precluded users from finding relevant material while others prolonged unproductive search. Three broad categories of misconceptions were identified. The first category includes misconceptions about the subject area itself. The second category includes misconceptions about the classification scheme. The third type of misconception is about the system's capabilities. In the remainder of this section, we describe these misconceptions, presenting examples for clarification.

4.1. Subject Area Misconceptions

A lack of expertise in the subject area leads to three related problems: not choosing appropriate terms to initiate a search, not having good a priori estimate about how much material there might be for a subject, and not expressing the query at an appropriate level of specificity.

4.1.1. Inappropriate Terms

The LCSH in the system consists of "official" terms. Users on the other hand, use terms they feel best express the "semantic content" of their problem. However, even though the user may be "close" in some sense to the official term, the term the user chooses may not yield any relevant material. To illustrate, one subject looking for books on "measure theory" used the term "measurement" instead. Only after more than 10 minutes browsing did he realize that the numerous citations (which had to do with all kinds of measurement such as pollution measurement, pollen count measurement, etc) had nothing to do with measure theory. In another case, a subject thought "information retrieval systems" was synonymous to "database management systems", which is not the case from the system's standpoint. Users detected such errors only after a significant amount of browsing.

4.1.2. No Estimate About the Volume of Relevant Work

Problems can also arise when a user does not have a good estimate about how much material exists in a subject area. A common misconception among users, particular Ph.D. students, is that their subject of interest is too specific for there to be books that are directly relevant. If an initial search attempt is unsuccessful, this bias tends to confirm the user's feeling that no relevant material exists. For example, on finding only one citation corresponding to the subject term "career", one subject thought that that was all the material to be expected -- not realizing the fact that there were over 50 citations listed under similar headings such as: "occupations", "professions" and "vocational guidance". Similarly, a subject was convinced that there was no book dealing with the "Contadora peace plan" because this topic is too specific and recent. Actually there were several books in the library that discuss the "Contadora peace plan" (which can be found simply by using the title search option!).

4.1.3. Expressing the Query at Inappropriate Level of Specificity

A common tendency is one of not expressing the query at the appropriate level of specificity. The use of a "broader entry first" strategy has been observed in other studies (Bates, 1977; Belkin et al., 1982). In this investigation, over 70% of our subjects used terms that were more general than they should have been. For example, one subject checked every citation under "statistics" when she actually needed something on "statistical power". In the other example, the subject browsed under "Nicaragua" and "Latin America" instead of the "Contadora peace plan". On the other hand, a minority of subjects used terms that were too specific, e.g. "Dempster-Shafer theory" and "software reusability" when they were really looking for literature on uncertainty and systems/software maintenance respectively. Such requests resulted in no matches. Because of the specific entry principle in LCSH, the broader entry first strategy often matched documents which were at the wrong level of specificity (Bates, 1986). One reason for why most queries tend to be expressed too generally appears to be one described in section 2, namely, that users often begin a search with an "anomalous state of knowledge". This tendency is probably reinforced by situations where prior experience with using specific terms results in no matches.

4.2. Classification Scheme Misconceptions

The backbone of the subject search option (SUB and SUBK) is the Library of Congress Subject Headings (LCSH) classification scheme. There were three types of misconceptions about this classification scheme: misinterpreting the terms used in subject headings, not realizing the indexing principles of LCSH, and simply ignoring the existence of LCSH.

4.2.1. Misinterpretation of Subject Headings

In contrast to situations where a user uses incorrect terms -- such as "measurement theory" instead of "measure theory", there are times when the user may in fact use correct terms, which nevertheless yield no matches. Reactions like the following were very common:

> "Human factors, no match? This is impossible

> "There should be something under organizational theory
> That is a standard area within organization."

This type of problem stems from lack of knowledge about the classification scheme, no cross-referencing facility within the system, and the incapability of the system to infer synonymous terms.

Finally there was some confusion between title and subject heading. Some users believed there should be an one-to-one correspondence between the title and the subject area. As one user remarked:

> "Invariably, the title is a reflection of its content.
> If someone has written something which has major contribution
> towards, let say, project management, there should be
> project management in the title. So if I use TIL or TILK, it
> works just as well as SUB."

This search strategy reflects a lack of knowledge about the classification scheme.

4.2.2. Not Realizing the LCSH Indexing Principles

A second problem was the lack of knowledge about the indexing principles of LCSH. There are three principles which were violated repeatedly: the specific entry principle, the whole document indexing principle, and the principle of subdivision. Based on the specific entry principle, subject headings assigned to a document are as specific as possible. However, users tend to think that a document classified under a certain heading should also be classified under a broader heading. The following remark is indicative of this type of misconception:

> "If there is something comes out of corporate planning,
> it should come out of planning too."

However, according to the specific entry principle, books classified under corporate planning would not be classified under planning.

Secondly, in order to reduce redundancy, LCSH indexers are trained to use a term that indexes a whole book, not a portion of it. This is referred to as the whole document indexing principle. Again, users without a knowledge of the LCSH violated this principle. The following remark illustrates this type of misconception:

> "Some books talked about several sub-topics in different chapters. Perhaps I can search from these sub-topics. The system should then suggest these books."

However, since a book is indexed for its entire content, a broader term can be assigned to a book than any of the topics it covers. For example, a book which covers queuing theory, linear programming, and inventory theory is likely to be assigned a subject heading like "operations research", and is therefore not accessible by its specific topics.

Lastly, most users were unaware of the subdivisions in LCSH. Standard subdivisions within subjects headings are features like *topical* (which limit a concept term to a sub-topic), *period* (time), and *local* (like geographic area). Without this knowledge, users tend to explore combinations of terms which for the most part are unproductive. For example, one subject spent about 10 minutes searching for information on the "Contadora peace plan" under the subdivisions of "Nicaragua" and "Latin America". Clearly, the "Contadora peace plan" is highly unlikely to be a standard subdivision given how specific it is (for such queries it makes more sense to use a keyword search based on title -- this would provide materials consisting of documents whose title includes the keyword in any position).

4.2.3. Not Consulting the LCSH

Only one out of the thirty subjects asked for or consulted the two volume LCSH handbooks. Faced with difficulty in generating system-recognizable terms, even experienced users did not consult it. If the user looking for information about "human factors" had looked at the LCSH handbook using the term "human factors" (an unofficial LCSH term), he would have found "human engineering" (an official LCSH term) via cross referencing.

We posit that the unwillingness of users to consult the two rather large looking LCSH handbooks stems from the infrequent usage of the system which might discourage the time investment needed to learn it. It is probably worthwhile to try to incorporate the LCSH knowledge into an online system.

4.3. System Misconceptions

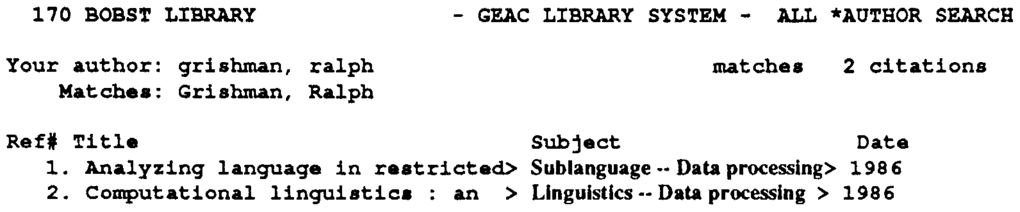
The last category of misconceptions is about the system itself. Twenty-five subjects exhibited erroneous perceptions about the system. System misconceptions are of two types: system's messages, and system's capabilities.

4.3.1. Misinterpretation of the System Messages

Some messages were misinterpreted by users due to a lack of precision in the system's language. For example, on typing "South Africa, sport" at the level where the system expected the input of a search option resulted the error message: "Your selection not recognized by the system." This message was interpreted by the user as stating that no document found under "South Africa, sport". In general, such misinterpretations arose because of the overly general content of the system's message.

Secondly, users confused the meaning of options. For example, some users were confused between the command PREVIOUS SCREEN which brought back the screen that was displayed previously, and BAC, which scrolled backward in the list of citations or headings. In other cases, users had difficulty distinguishing between the command IND which displayed the list of subject headings and the command CIT which displayed the list of matched documents.

Users also tended to ignore vital information on the screen. In many situations, users were not aware of subject headings printed on the screen which were relevant to the query. These subject headings were either displayed along with the titles of the documents (see the terms listed under "Subject" in Figure 3) or as part of a detailed description of a book (see the terms listed after "SUBJECTS" in Figure 4). We posit that this occurred because too much incidental information was presented on the screen, causing the user to overlook the important cues.

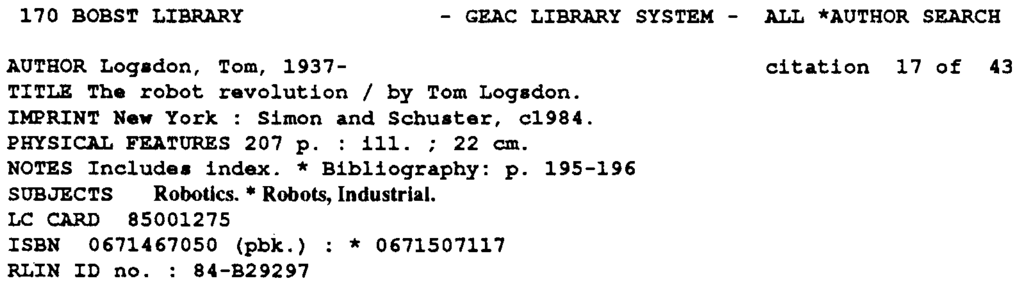


```
 170 BOBST LIBRARY              - GEAC LIBRARY SYSTEM -  ALL *AUTHOR SEARCH

Your author: grishman, ralph                        matches  2 citations
    Matches: Grishman, Ralph

Ref# Title                         Subject                   Date
   1. Analyzing language in restricted> Sublanguage -- Data processing> 1986
   2. Computational linguistics : an  > Linguistics -- Data processing > 1986
```

FIG. 3. The screen display of matched titles



```
 170 BOBST LIBRARY              - GEAC LIBRARY SYSTEM -  ALL *AUTHOR SEARCH

 AUTHOR Logsdon, Tom, 1937-                         citation  17 of  43
 TITLE The robot revolution / by Tom Logsdon.
 IMPRINT New York : Simon and Schuster, c1984.
 PHYSICAL FEATURES 207 p. : ill. ; 22 cm.
 NOTES Includes index. * Bibliography: p. 195-196
 SUBJECTS    Robotics. * Robots, Industrial.
 LC CARD  85001275
 ISBN  0671467050 (pbk.) : * 0671507117
 RLIN ID no. : 84-B29297
```

FIG. 4. The screen disolav of full citation information

4.3.2. Not Understanding the System Capabilities

There were three misconceptions about the system's capabilities. The first was the lack of clear understanding of the capability of each of the seven search options. Confusion was observed among the SUB, SUBK, TIL, and TILK options. In reality, SUB can only search for the subject headings if the user's terms happen to be in the leftmost position of an official LCSH subject heading. For example, "economic" will match headings like "economic development" but not headings like "international economic relations". SUBK on the other hand takes only one word but matches all headings which have that word appearing in them, regardless of position. For example, "business planning" can be matched by using "planning" as the keyword in SUBK. The same rules apply to TIL and TILK. Several subjects actually used these options interchangeably or did not realize the potential fruitfulness of using one option over others. This misconception reflects a limited knowledge of the system's functionality and a low frequency of use.

A second major problem related to the system's capabilities was a lack of understanding of the system's match/search method. The system finds documents by matching alphabetically the terms supplied by the user. Six subjects exhibited lack of understanding of this process. For example, one subject typed in "South Africa, sport" after he had already browsed unsuccessfully all headings that had matched "South Africa". In another case, a subject typed in "salt substitute" which matched alphabetically close headings starting from "salt -- social aspects". She didn't realize there were subject headings before the first screen of the list of headings displayed in front of her that also had to do with other aspects of salt. Typing BAC (to move backward in the list of matched headings), would have shown "salt -- physiological effect", which she was looking for. Finally, the use of operators such as "AND" and "OR" in non-Boolean search seems to suggest that users expect certain set operations to be performed automatically by the system. An example of this type of misconception is apparent in the following subject search (SUB) query: "Dempster-Shafer theory and expert systems" where the user expected the "AND" to be interpreted by the system to perform set intersection.

The last type of problem resulted from a lack of knowledge about "levels" of menus in the system. Basically the system's menus are at several levels as shown in Figure 5. Users had problems remembering their position during their interaction. For example, instead of using CIT to return to scanning a citation list, many users typed in IND which incorrectly brought them back to the list of subject headings (one level higher than CIT), from where they again had to begin looking for the individual citations within headings. For example, a subject initially matched 107 citations using the term "game theory". After reviewing the detailed citation of a book in this list, instead of returning to the original position in the citation list (popping up one level), he went into the outermost system loop -- restarting the search by choosing SUB, "game theory", and moving forward in the citation list. This process was repeated 11 times. Apart from limited knowledge about how to traverse the system menus, this type of wasteful search is probably the consequence of limited human short term memory. In the absence of an indication of the level of the dialog from a system, the user often tends to go back to the top level unnecessarily.

```
  selection of search options (see Figure 2, CAT to return)
    list of subject headings (IND to return)
       list of citations (CIT to return)
 brief description of the citation (BRF to return)
detailed description of the citation (see Figure 4, FUL)
```

FIG. 5. The levels of menus

Figure 6 summarizes the hierarchy of misconceptions described above and the causes of misconceptions. Connections from the causes to misconceptions reflect our hypotheses for the misconceptions described in this section. We have not drawn physically the actual linkages since there are too many such connections, which would congest the figure. Instead, the causes for each misconception are indicated in the square bracket listed with each misconception. The numbers in the square bracket correspond to causes in the lower part of the figure.

5. Representation of Misconceptions

In this section, we will present a knowledge representation scheme which helps account for the user misconceptions about the subject area, the classification scheme, and the system. Our ultimate goal is to augment the retrieval system with the "correct" representation so that it might be able to compare the two and take steps to alleviate the misconceptions. We describe both types of representations in this section.

5.1. Erroneous and Correct Representations

In this subsection, we will examine each type of misconception shown in Figure 6 in detail. Examples of the erroneous user representations along with the correct representations are provided.

A Taxonomy of Misconceptions:

1. **Subject Area Misconceptions**
   - 1.1. Misinterpretation of the terms [1.1]
   - 1.2. Inappropriate level of specificity [1.1/1.5]
   - 1.3. Not realizing the amount of relevant works [1.1]

2. **Classification Scheme Misconceptions**
   - o Misinterpretation of subject headings
     - 2.1. Treating title as subject [1.2]
     - 2.2. Using unofficial terms [1.2/2.5/2.6]
   - o Not realizing the LCSH indexing principles
     - 2.3. specific entry [1.2]
     - 2.4. whole document indexing [1.2]
     - 2.5. subdivisions [1.2]
   - o Not consulting LCSH [1.2/1.6]

3. **System Misconceptions**
   - o Misinterpretation of the system message/display
     - 3.1. misinterpretation of the system message [2.1]
     - 3.2. confusion about the system options [2.1/2.2]
     - 3.3. missing vital information on the screen [1.7/2.4]
   - o Not understanding the system capabilities
     - 3.4. not understanding top-level search options [1.3/1.6/2.2]
     - 3.5. not understanding the match method of the system [1.3]
     - 3.6. not understanding the levels of menus [1.3/1.4/2.3]

Causes of Misconceptions:

1. **User-Attributed Causes**
   - o knowledge based characteristics
     - 1.1. subject knowledge inadequacy
     - 1.2. low classification scheme knowledge
     - 1.3. low knowledge of system functionality
   - o general human characteristics
     - 1.4. limited short term memory
     - 1.5. prior experience in using the system
     - 1.6. indifference due to low frequency of usage
     - 1.7. carelessness
2. **System-Attributed Causes**
   - o poor general system features
     - 2.1. ambiguous messages
     - 2.2. poor help screen
     - 2.3. non-transparent structure
     - 2.4. non-highlighting vital information
   - o problematic classification scheme
     - 2.5. no cross-referencing of terms
     - 2.6. no semantics in matching

FIG. 6. Users' misconceptions and the causes

5.1.1. Representation for Subject Area Misconceptions

We can think of the user's knowledge about the subject area as a large semantic network with nodes representing terms, and links representing the relationships between terms. Users also estimate the number of citations classified under different terms. This can be thought of as an attribute of a term.

As described earlier in Section 4, we observed three types of subject area misconceptions: misinterpretation of the terms, inappropriate level of specificity, and not realizing the amount of relevant work in the field. Each of them can be conceptualized in terms of a semantic network.

1. **Misinterpretation of terms:** This refers to an inappropriate meaning being assigned to a term (or a wrong term being used). In this situation, the user assumes (mistakenly) a link between unrelated terms. For example, a subject treated "measurement theory" as a sub-topic of "statistical theory" as shown in the left portion of Example A in Figure 7. The correct representation takes "measure theory" instead of "measurement theory" as a term narrower than "statistical theory" (see the right portion of Example A in Figure 7).

2. **Inappropriate level of specificity:** Users who express their queries too broadly tend to assume (sometimes mistakenly) that a specific term is not recognizable by the system. For example, a subject did not believe there could be a term as specific as "Contadora peace plan" (see Example B in Figure 7) when in fact the term "Contadora" would have been recognized using the subject keyword search. Conversely, users may specify a term which is too narrow for the system to recognize.

3. **Not realizing the amount of relevant work:** As described earlier, users may have no good estimate of the number of citations (an attribute of a term as will be discussed in Section 5.2.1) that will match a search term. In Example C (Figure 7), this is shown as a "?". If the user feels that the term is too narrow to yield any matches (i.e. matches: 0), he may not even attempt to use the (correct) term, as was the case in the example involving "statistical power".

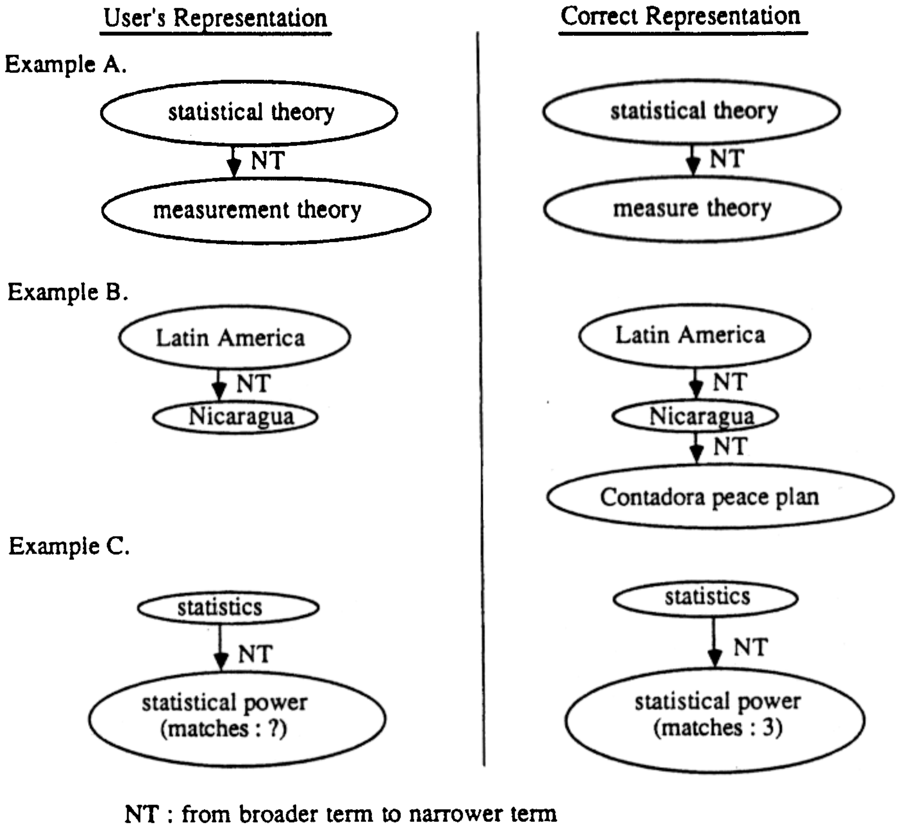


FIG. 7. Examples of representation of subject area misconceptions

5.1.2. Representation for Classification Scheme Misconceptions

Users have different classification scheme misconceptions, including: lack of knowledge of the specific entry principle, whole document indexing principle, subdivision principle used by the LCSH and ignorance about subject headings. Such misconceptions can be analyzed in terms of a semantic network and production rules.

1. **Misinterpretation of subject headings:** Users have a tendency to treat unofficial terms as subject headings. For example, a subject used "human factors" for "human engineering" (see left portion of Example A in Figure 8). The retrieval system, however, can only recognize "human engineering", treating "human factors" as synonymous unofficial term (see right portion of Example A in Figure 8).
2. **Lack of knowledge of the specific entry principle:** Even though users may have correct knowledge about the terms and the relationships between them, they may think (erroneously) that citations that match a term will also be listed under a broader term (represented using production rule <X contains Y> as appeared in Example B of Figure 8), while the LCSH's specific entry principle prevents this from happening for reasons described in Section 2 (X and Y are actually disjoint).
3. **Lack of knowledge of the whole document indexing principle:** Several users believed that citations that match a topic should also be listed under all its sub-topics (described in the previous section). This is shown as the production rule listed in the left portion of Example C in Figure 8. LCSH, however, classifies citations which cover several sub-topics under the broader topic only, not under each individual sub-topic (see the production rule in the right portion of Example C).
4. **Lack of knowledge of the subdivision principle:** Users who have no knowledge about the subdivision principle are generally ignorant of the existence of certain subject headings. This leads to missing nodes and links in the semantic network. For example, a novice user looking for finance information in France is not likely to use the regional subdivision such as "finance -- France" (the missing node). This example is shown in Example D of Figure 8.

5.1.3. Representation for System Misconceptions

Users appear to have widely varying perceptions about system messages and its matching methods. It is therefore difficult to represent this type of misconception. However, the other two system misconceptions, namely, misconceptions about search options and the levels of menus, are easy to capture.

1. **Not understanding the search options (SUB = SUBK = TIL = TILK):** Searchers tend to have problems in distinguishing between search using subject headings (SUB), keyword search using subject headings (SUBK), title search (TIL), and keyword search using title (TILK). Due to users' reluctance to learn about the system, this misconception about search options is hard to remove. However, in the next section we will present heuristics for dealing with this problem.
2. **Not understanding the levels of menus (CAT = IND = CIT):** As illustrated in Figure 5, Bobcat has different levels of menus. Users, in interacting with the system, had problems remembering and distinguishing between the top three levels of menus, namely, CAT (to return to the top level search option), IND (to return to a list of subject headings), and CIT (to return to a list of matched citations). They generally fell into cycles and/or performed redundant operations.

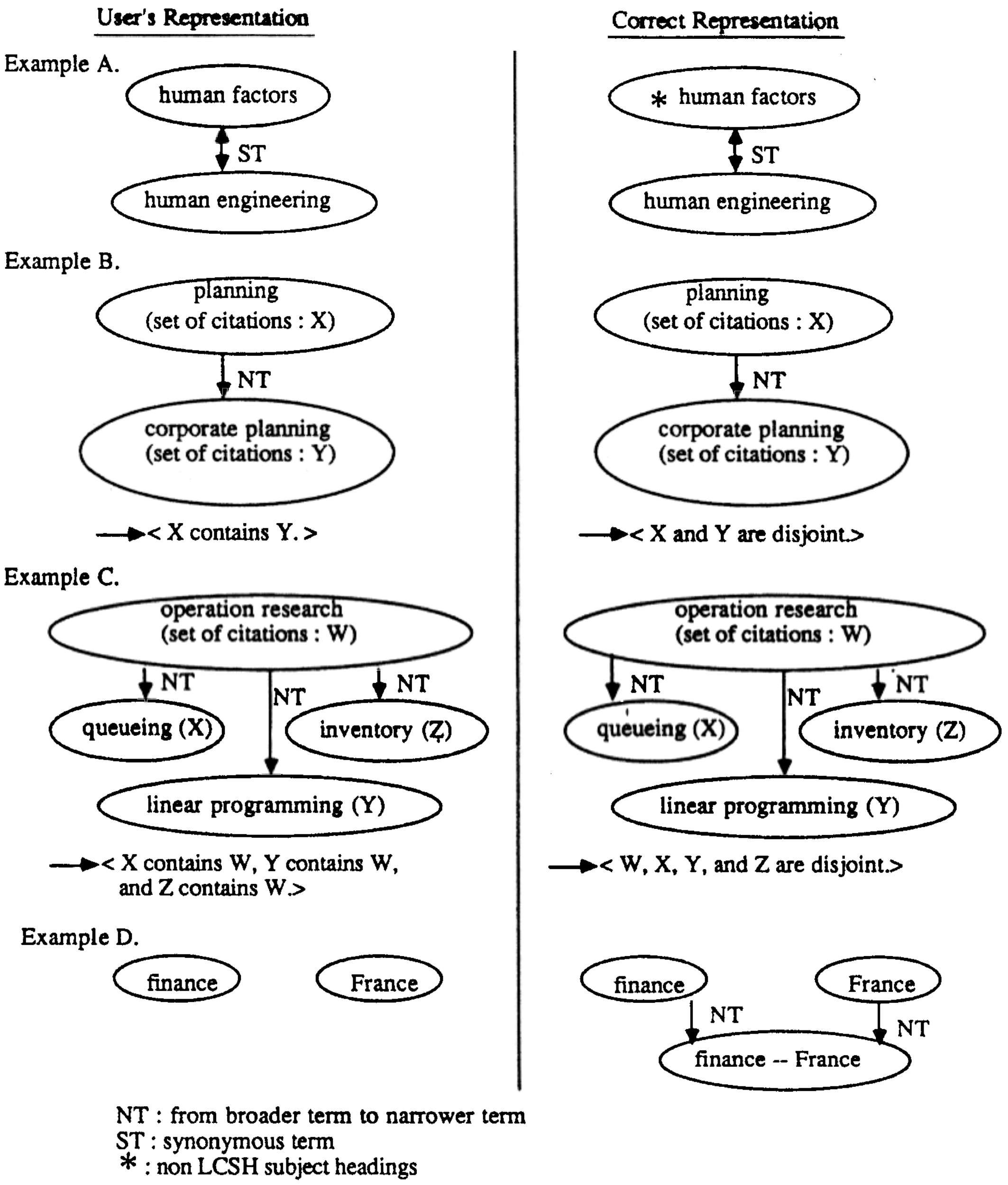


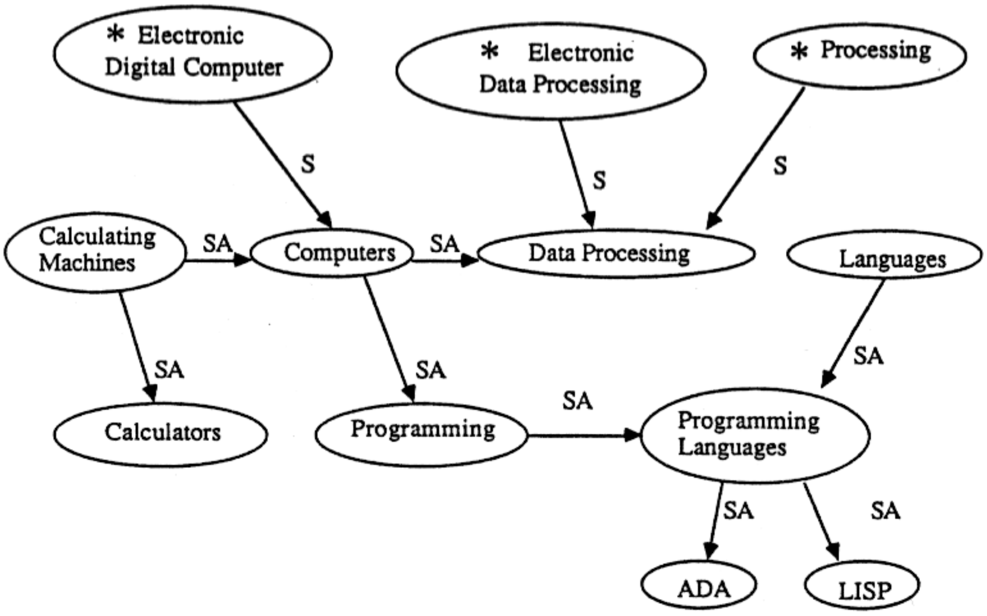
FIG. 8. Examples of representation of classification scheme misconceptions

5.2. Alleviating Misconceptions

In this subsection, we propose a knowledge representation scheme and heuristics for alleviating the various types of misconceptions.

5.2.1. Frame-based Semantic Network for Representing Knowledge

A semantic network structure is appropriate for constructing an online thesaurus based on subject area knowledge and classification scheme knowledge. The existing thesaurus (hardcopy) contains not only official terms (subject headings), but also "unofficial" terms which point to the official terms. The thesaurus can be viewed as a large semantic network of terms (concepts) where links are of two types: relations between unofficial and official terms, and set-superset relationships (like IS-A links). Figure 9 shows a portion of the semantic network corresponding to the LCSH classification scheme. Terms and the cross referencing structure of the LCSH are incorporated[1]. By including the knowledge (terms and relationships between terms) from other thesauri, the scope of the online thesaurus can be extended.



*	--- unofficial term
S	--- see reference (lead unofficial term to official term)
SA	--- see also reference (lead broader term to narrower term)

FIG. 9. A sample network of LCSH terms

[1] The S links in Figure 9 are equivalent to the ST links in Figure 8. The SA links are equivalent to the NT links in Figure 7 and 8.

In order to make the thesaurus useful for alleviating user misconceptions, it is necessary to specify more extensively, the semantics of the network. Specifically, the following slot values (with an example in Figure 10) are necessary for each term:

1. Term: Specify the name of the term.
2. Type of term: Specify whether the term is an official subject heading or an unofficial term.
3. Definition: Provide a brief definition for the term.
4. SA: Specify all narrower terms.
5. XX: Specify all broader terms. The SA links and XX links represent the relationships between official terms.
6. S: Specify the synonymous official terms. This link relates unofficial terms and official terms.
7. X: Specify the synonymous unofficial terms.
8. Number of matched citations: Specify the number of citations indexed under the official term. It can be computed when the user requests. This information can help users determine the amount of relevant work associated with the subject heading.
9. Titles of matched citations: Specify the titles of matched citations. Users can examine this list of titles to check the relevance of citations to their queries. This, again, is computed when user requests.

The frame-based semantic network which includes subject area and classification scheme knowledge can be used in several ways. First, users can actively interact with the online thesaurus. By examining the definitions of terms, checking the number and titles of matched citations, and traversing the network via the links (SA, XX, S, and X), users can examine term meanings and relationships among terms. Secondly, the system can actively assist in the search process by automatically performing a "terms translation" process along the S links to generate the official subject headings whenever necessary. The system can also perform a "spreading activation" along the SA or XX links to identify the most appropriate search terms from the original search terms (i.e. using path intersections), in effect performing a process similar to query refinement (Chen & Dhar, 1987). Designing the algorithms for achieving these types of functionality is the current phase of this research.

5.2.2. Heuristics for Improving Search Performance

We propose several heuristics which should help users avoid many of the obstacles we observed. These heuristics fall into two categories: rules for online browsing and rules for selecting appropriate search options.

Browsing Rules:

1. Look for the subject headings of each citation. They can be used as search terms.
2. Browse both forward and backward in the list of headings when a subject heading produces matches. Relevant subject headings may be in the previous or subsequent screens.

```
Term Object Frame:

     {Term: (name of the term)
           Type of term: (* for unofficial term; nil for official term)
           Definition: (textual explanation)
           SA: (list of narrower terms)
           XX: (list of broader terms)
           S: (list of synonymous official terms)
           X: (list of synonymous unofficial terms)
           Number of citations: (integer)
           Titles of citations: (list of matched titles)
     }

   Example of Term Instance:

     {Term: computers
           Type of term: nil
           Definition: "A programmable electronic device that can store,
                        process, and retrieve data."
           SA: (data-processing programming)
           XX: (calculating-machine)
           S: nil
           X: (electronic-digital-computer)
           Number of citations: 56
           Titles of citations: ("Introduction to Computers" "Computers'
                                  ...)
      }
```

FIG. 10. Frame-based representation for alleviating misconceptions

3. If the number of matched subject headings or citations is too large, and little or no relevance is observed after browsing a screenful, make the search terms more specific.

Search Options Rules:

1. For subject areas that can be described using one word, use keyword searches (TILK and SUBK); otherwise use the standard title or subject search options (TIL and SUB). This increases the likelihood of a match since a keyword can match in any position.
2. If possible, begin by performing a known-item search first. Use author or title search (AUT, AUTK, TIL, and TILK) to identify one or a few citations and examine subject heading information from the detailed citation. More citations can be obtained by performing subject search (SUB) using the derived subject headings.
3. If the documents matched are too few, try keyword search options (TILK, AUTK, and SUBK) to broaden the search space.
4. For terms which are likely to appear in the titles of books, perform the title-based searches (TIL and TILK) instead of the subject searches (SUB and SUBK).

These rules should enable users to make the most of the capabilities provided by existing retrieval systems.

6. Discussion

In previous research (Chen & Dhar, 1987) we identified an important component of the librarian/user consultation model as one where on the basis of cues, the librarian stereotypes the user. Specifically, hypotheses about a user's information needs are based on the purpose of search and the user's level of education. The advantages of stereotypical user modeling in question answering systems has been demonstrated in (Rich, 1979; Rich, 1983).

The taxonomy of misconceptions and the frame-based semantic network representation we have described in this paper provide a sound basis for building an intelligent user modeling component within an information retrieval system. Information about the user and his query can be obtained through a few questions and from an analysis of the discourse. The system should be able to use such information to create an individual, implicitly-inferred model for a user (Rich, 1983; Borgman, 1988). The user model we have in mind would be similar to the student model constructed in Intelligent Computer-Aided Instruction (ICAI) systems (Dede, 1986; Woolf & McDonald, 1984; Sleeman & Brown, 1982).

The output of the user-modeling of the system could be used in two ways. It could be used to plan an appropriate search sequence, that is, suggest what search options to use and how. As we described earlier, known-item searches can be used to anchor and/or determine terms that can be used to perform a focused subject search. The output of the user-modeling component could also be fed into a terms translation process which would require knowledge about the subject area integrated with the classification scheme into an "online thesaurus" based on the frame-based semantic network knowledge representation scheme we have proposed. Armed with the appropriate search heuristics and the subject area and classification scheme knowledge, a retrieval system could become a responsive and intelligent information intermediary capable of providing interactive support to a variety of users. Based on the results of this and the earlier study (Chen & Dhar, 1987), we have been designing a prototype that is nearing completion.

References

ASHBY, W. R. (1973). *An Introduction to Cybernetics*, London: Methuen.

BATES, M. J. (1977). Systems meets user: problems in matching subject search terms. *Information Processing and Management*, 13, 367-368.

BATES, M. J. (1979). Information search tactics. *Journal of the American Society for Information Science*, 30, 205-214.

Bates, M. J. (1986). Subject access in online catalog: a design model. *Journal of the American Society of Information Science*, 37, 357-376.

BELKIN, N. J., ODDY, R. N. and BROOKS, H. M. (1982). Ask for the information retrieval: Part I. background and theory. *Journal of Documentation*, 38, 61-71.

BORGMAN, C. (1988). Retrieval systems for the information seeker: can the role of the intermediary be automated? *Panel Session, CHI'88 Conference*

Proceedings, Addison Wesley.

CHAN, L. M. (1986). *Library of Congress Subject Headings: Principles and Application*, Littleton, CO: Libraries Unlimited.

CHEN, H. & DHAR, V. (1987). Reducing indeterminism in consultation: a cognitive model of user/librarian interaction. *Proceedings of the National Conference on Artificial Intelligence*.

COCHRANE, P. A. & MARKEY, K. (1985). Preparing for the use of classification in online cataloging systems and in online catalogs *Information Technology and Libraries*, 4, 91-111.

DEDE, C. (1986). A review and synthesis of recent research in intelligent computer-assisted instruction. *International Journal of Man-Machine Studies*, 24, 329-353.

FURNAS, G. W., LANDAUER, T. K., GOMEZ, L. M. and DUMAIS, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30, 964-971.

GOMEZ, L. M. and LOCHBAUM, C. C. (1984). People can retrieve more objects with enriched keyword vocabularies. But is there a human performance cost? In B. SHACKEL, Ed. *Human-Computer Interaction--Interact '84*. North-Holland, Amsterdam, 257-261.

GOOD, M. D., WHITESIDE, J. A., WIXON, D. R. and JONES, S. J. (1984). Building a user-derived interface. *Communications of the ACM*, 27, 1032-1042.

HJERPPE, R. (1985). Project HYPERCATalog: visions and preliminary conceptions of an extended and enhanced catalog. *Proceedings of IRFIS*, 6th, Frascati, Italy, pp15-18.

JACOBY, J. and SLAMECKA, V. (1962). *Indexer consistency under minimal conditions*. Bethesda, MD: Documentation, Inc..

KASKE, N. K. and SANDERS, N. P. (1980). On-line subject access: the human side of the problem. *RQ*, 20, 52-58.

LANCASTER, F. W (1979). *Information retrieval systems*. John Wiley & Sons, Inc..

LARSON, R. R. and GRAHAM, V. (1983). Monitoring and evaluating MELVYL *Information Technology and Libraries*, 2, 93-104.

MARKEY, K. (1981). Levels of question formulation in negotiation of information need during the online presearch interview: a proposed model. *Information Processing & Management*, 17, 215-225.

MARKEY, K. (1984). *Subject Searching in Library Catalogs: Before and After the Introduction of Online Catalogs*. Dublin, OH: OCLC.

MATTHEWS, J. R., LAWRENCE, G. S. and FERGUSON D. K. (1983). *Using online catalogs: a nationwide survey*. New York: Neal-Schuman.

NOERR, P. L., BIVINS NOERR, K. T. (1985). Browse and navigate: an advance in database access method. *Information Processing & Management*, 21, 205-213.

RICH, E. (1979). User modeling via stereotypes. *Cognitive Science*, 3, 329-354.

RICH, E. (1983). User are individuals: individualizing user models *International Journal of Man-Machine Studies*, 18, 199-214.

ROCHELL, C. C. (1984). *A study of user success with an online catalog*. New York: The Libraries.

SHOVAL, P. (1985). Principles, procedures and rules in an expert system for information retrieval. *Information Processing & Management*, 21, 475-487.

SLEEMAN, D. H. & BROWN, J. S. (1982). Introduction: intelligent tutoring systems. In *Intelligent Tutoring Systems*, Academic Press, pp. 1-8

STEVENS, M. E. (1965). *Automatic Indexing: A State-of-the-art Report*. Washington, DC: U.S. Government Printing Office.

TAGLIACOZZO, R. & KOCHEN, M. (1970). Information-seeking behavior of catalog users. *Information Storage & Retrieval*, 6, 363-381.

TAYLOR, R. S. (1968). Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29, 61-71.

TOU, F. N., WILLIAMS, M. D., FIKES, R., HENDERSON, A. and MALONE, T. (1982). RABBIT: an intelligent database assistant. *Proceedings of the National Conference on Artificial Intelligence*.

WILLIAMS, M. D. (1984). "What makes RABBIT run ?". *International Journal of Man-Machine Studies*, 21, 333-352.

WOOLF, B. & McDONALD, D. D. (1984). Building a computer tutor: design issues. *Computer*, 17, 61-73.