Interaction in Information Retrieval: Selection and Effectiveness of Search Terms

Amanda Spink*
*School of Library and Information Sciences, University of North Texas, P.O. Box 13796,
Denton, TX 76203-3796. E-mail: spink@lis.admin.unt.edu*

Tefko Saracevic
*School of Communication, Information and Library Studies, Rutgers University, 4 Huntington St.,
New Brunswick, NJ 08903. E-mail: tefko@scils.rutgers.edu*

We investigated the sources and effectiveness of search terms used during mediated on-line searching under real-life (as opposed to laboratory) circumstances. A stratified model of information retrieval (IR) interaction served as a framework for the analysis. For the analysis, we used the on-line transaction logs, videotapes, and transcribed dialogue of the presearch and on-line interaction between 40 users and 4 professional intermediaries. Each user provided one question and interacted with one of the four intermediaries. Searching was done using DIALOG. Five sources of search terms were identified: (1) the users' written question statements, (2) terms derived from users' domain knowledge during the interaction, (3) terms extracted from retrieved items as relevance feedback, (4) database thesaurus, and (5) terms derived by intermediaries during the interaction. Distribution, retrieval effectiveness, transition sequences, and correlation of search terms from different sources were investigated. Search terms from users' written question statements and term relevance feedback were the most productive sources of terms contributing to the retrieval of items judged relevant by users. Implications of the findings are discussed.

Introduction

Information retrieval (IR) systems emerged in the 50s and 60s as static, batch processing systems. Starting in the 70s, with the revolutionary symbiosis between computer and communication technologies, the access to IR systems became dynamic and interactive. In practice, interaction became *the* most important feature of information retrieval. Means, ways, models, and types of IR interactions are still evolving, changing, and, at times, improving. However, we still do not fully understand the many complex aspects of interactive processes, despite a number of theoretical and experimental studies and scholarly treatises (Ingwerson, 1992, 1996). Furthermore, most of the research and development in IR, concentrating on the improvement of effectiveness in automatic representation and searching, has treated IR systems and processes as static and not as dynamic or interactive (Saracevic, 1995). Such research, conducted now for 30 years, has reached a certain maturity, as evidenced by the Text Retrieval Conference (TREC) experiments (Harman, 1995). In contrast, research on the interactive aspects of IR has not reached maturity; it may be said to be emerging out of infancy. There is a clear need for two things: to concentrate more research in IR on interactions, to more resemble what is actually going on in practice, and to attempt to use what was found in interactive research for the design and development of improved IR interfaces and processes. Because we know so relatively little about the complexity of various interactive variables, and particularly about their effects, the design of IR interfaces ''is not straightforward'' (Belkin & Croft, 1992).

In this paper we concentrate on the *selection of search terms,* one of the key objectives and processes of IR interaction. We chose to study a complex, two-part interaction process where the selection of search terms takes place: the interaction between a user and an intermediary, before and during on-line searching, and the interaction between the user and intermediary on the one side and the IR system on the other side. In other words, we chose a mediated on-line IR interaction, to discern the sources of search terms and then to observe the effectiveness of terms from each source. We made observations from a particular type of interactive process, as found in mediated IR practice. While we fully realize that, strictly speaking,

* To whom all correspondence should be addressed.

Received April 20, 1995; revised May 21, 1996; accepted October 2, 1996.

our conclusions pertain only to the interaction type or model chosen and to the data from this case study, we also believe that they provide an illumination of the interactive IR process in general.

Which search terms should be selected for a given query to represent a user's information problem? This is a key issue and problem in searching of IR systems in both practice and research (Fidel, 1991). By extension, this also involves the problem of sources: *Where should search terms be selected from?,* as well as the problem of the dynamics of selection: *What interactive processes can aid in selection?* Search terms are a central determinant in IR, and selection of search terminology is a driving force and variable in IR processes (Saracevic, Mokros, Su, & Spink, 1991). In themselves, these problems define the importance of any and all investigations about search terms.

Objectives

The results reported here stem from a study that is a long-term effort involving a number of projects by the same and expanding group of investigators whose collective aim is to contribute to the formal characterization and better understanding of elements and processes involved in human information seeking and retrieving and in IR interactions, from the human perspective. A number of papers and dissertations are related to the analysis of the same or similar data from different perspectives (Mokros, Mullins & Saracevic, 1995; Saracevic, Mokros, & Su, 1990; Saracevic et al., 1991; Spink, 1993a,b; Wu, 1992). A detailed description of the data collection, models and approach can be found in those papers, thus only a summary will be given here.

The objectives of the particular analysis reported here is to use the data from the larger study and to concentrate on the investigation of sources of search terms, the retrieval effectiveness of terms from different sources, and the dynamics of search term selection, as observed during real-life interactions between intermediaries, users and a given on-line system. By ''real-life'' we mean that the study involved: real users, with real questions; interaction with professional intermediaries; searching using a commercial on-line vendor (DIALOG); and the usual premises and policies of the institution involved (Rutgers University libraries). In other words, this was a naturalistic case study and not a laboratory study, with all its strengths and weaknesses.

Given the model of interaction developed for the study as a whole (described briefly below), the approach to the analysis of search terms involved:

(1) Classifying the sources of terms.
(2) Identifying the sources of those terms that retrieved items judged relevant by users.
(3) Observing the transitions in sequences of search term use.
(4) Examining the correlation between terms from different sources and search outcomes.

Several publications from the project have already dealt with partial results about search terms, incorporating differing aspects of the process of search term selection (Saracevic, et al., 1991; Spink, 1993a, 1994, 1995; Spink & Saracevic, 1992a,b, 1993a,b). In this paper, we unify those observations, with further analysis to cover all four aspects listed above.

Approaches to Study of Search Term Selection

Two approaches, algorithmic and human, are currently used to examine the process of search term selection in IR. The two approaches are quite distinct, different, and even isolated from each other. As research proceeds on both fronts, the basis for meaningful interaction between researchers from different approaches increases.

Algorithmic approach

In general, this approach uses the text of a question (topic) and/or a document found relevant by the user, as input for an automatic technique to select search terms, with logical connectors, and/or term weights. Some techniques use statistical approaches for both documents and questions; they select weighted terms for a search query from the question. Examples of recent efforts along these lines can be found in the large evaluative project TREC (Harman, 1995). Another approach is the automatic relevance feedback technique; here the concentration is on using search terms from documents found relevant. Various relevance feedback algorithms have been evaluated as to their effectiveness (Efthimiadis, 1993; Rocchio, 1971; Salton & Buckley, 1990; Spink & Losee, 1996), but it is not yet clear which are most effective in given contexts. Natural language processing techniques (stemming, lexicon-based normalization, phrase formation and augmentation, context-based clustering, and others) have also been used for query expansion and the selection of search terms (Strzalkowski, 1995) but with indeterminate results. Some of this research has also focused on the development of ''intelligent'' computer-based intermediaries, interfaces, and expert systems to help end-users in the selection of search terms and/or in the adjustment of terms during database searching (Croft, 1987). In any case, the selection of search terms for automatic query formulation and/or expansion, using statistical or linguistic models as theoretical base, continues to be a major area of IR research.

Unfortunately, interaction was *not* largely involved in this rich body of research, including the investigation of relevance feedback techniques. Even the massive effort of TREC ‘‘in its present form is antiphatic to interactive information retrieval’’ (Robertson, Walker, & Hancock-Beaulieu, 1995). Theirs was the first effort to include interactivity in IR tests in general and in TREC in particular. TREC 4 (conducted in October 1995) contained 11 projects involving interaction; inevitably as this effort progresses, problems related to the selection of search terms, and the help provided to users for selection/elaboration, will become a major component of research.

Human approach

Research within this approach stresses the investigation or observation of human decision-making and behavior and cognitive variables during the process of interactive information retrieval. The approach derives its theoretical base mostly from cognitive science and psychology. Typically, the human approach examines and models the behavior of search intermediaries and end-users (Belkin, 1984; Fidel, 1991). It uses whatever a user provides to represent an information problem and the user’s experiences and context to derive and modify a set of search terms. A user may select search terms alone or in mediation with a human search intermediary. During mediated database searching, search term selection can be *dyadic* between a user and intermediary, where the intermediary on his/her own interacts with an IR system, or *triadic* between a user, intermediary, and an IR system are together, i.e., where the user is present before and during the interaction with the IR system.

In the project as a whole, we chose to observe and collect evidence about interactions involving users, intermediaries, and an IR system; that is, we have chosen the context of the triadic interaction. A question can be raised: *Because end-user searching is rising explosively, why study mediated on-line retrieval, which is a vanishing breed of service?* The answer is simple. As in other situations where user modeling or diagnosis are involved, we do not understand the process well. That is, as yet, we simply do not understand well the interactive processes in IR in all their complexity at different levels. As a result, user modeling, although discussed at great length, is not incorporated in to current IR systems design, and the ‘‘intelligent’’ interfaces have as yet not an ounce of intelligence when it comes to such modeling. If we wish to enhance user modeling and incorporate it with any degree of success in the design of IR systems, then we must study and understand first what is going on in interactions involving human intermediaries in particular.

Related Studies

IR literature contains many book chapters and articles devoted to the selection of search terms. Most of this literature is descriptive or prescriptive (Blair, 1990; Harter, 1986). However, many studies have examined various aspects of search term selection. One of the most popular topics was contrasting the retrieval performance of free-text terms on the one hand, with descriptors from thesauri (controlled vocabulary terms), on the other hand (Bates, 1988; Blair & Maron, 1985; Cleverdon, 1967; Dubois, 1987; Fidel, 1986, 1991; Keen, 1973; Lancaster, 1980; Markey et al., 1982; Parker, 1971; Rowley, 1994). Note that we have also addressed the selection of terms from a thesaurus as a source but strictly limited to the type of interaction and analyses we have chosen. Thus, we have *not* addressed the effectiveness of free versus controlled vocabulary as in some of the mentioned studies.

A growing body of studies has investigated the selection of search terms with data from real search situations. Here are several studies that particularly relate to our study:

- In a large research study of search terminology, Fidel (1991) analyzed 281 on-line searches by 47 searchers. She derived a decision-tree routine for the selection of search terms, contrasting controlled versus free-text searching, and defining several searching styles. She noted that both types of search terms, text words and controlled vocabulary terms, complement each other, evenly splitting in numbers (50–50) for databases with a thesaurus. She was also concerned with *moves*—modifications of search strategies that are aimed at improving the results of the search, as we are.
- Saracevic, Kantor, Chamis, and Trivison (1988) compared the search terms selected by experienced searchers for the same information problem. They found considerable difference in search terms selected for the same questions by different professional searchers: the average overlap of terms selected by five searchers for the same questions was 27%.
- Hsieh-Yee (1993) found that novice and experienced searchers differ in their selection and manipulation of search terms for the same information problem, including their use of synonyms. She found that novices relied on nonthesaurus search terms and used fewer numbers of sources than more experienced searchers.
- Bates, Wilde, and Siegfried (1993) found that search terms used by humanities scholars differed markedly from the types of search terms used by physical scientists. Humanities scholars selected more chronological terms, geographical terms, names of works, and individual names than subject terms. Searches by physical scientists used mainly subject terms or common terms. Bates and coworkers concluded that ‘‘searches in the humanities may be inherently more complex than in the sciences.’’

The study reported in this article follows these examples. We suggest, as did all of the aforementioned authors, that findings regarding the selection and effectiveness of search terms from different sources can provide guidelines for database searching practice and also for design of interfaces and/or algorithms to help users with search term selection and query reformulation, thus possibly linking the human and algorithmic approaches in IR.

Framework: A Stratified Model of IR Interaction

Various models of IR processes and systems reflect interactive aspects in different ways. The traditional model, used in most algorithmic approaches reviewed above, represents IR as a two-pronged set (system and user) of elements and processes converging on matching (for diagram and description, see Belkin & Croft, 1992). In the traditional model, the system prong involves documents/texts, that were represented in a given way and then organized in to a file, and ready for the matching process. The user prong starts with a user's information problem/need, that is represented (verbalized) by a question, which is transformed into a query acceptable to the system, and then matching between the two representations (texts and query) occurs. A feedback function is included that allows for the modification of representations but usually involves only the modification in the user prong. The strength of the model is that it allows for the straight forward isolation of variables and for comparison. But, the model has often been criticized for weaknesses (Belkin, 1993). Primarily, interaction is not directly depicted in the traditional model at all. It is assumed under feedback. In turn, feedback was treated mostly as an instrument for query modification. Yet, even a most casual observance of IR interaction can see that there is much more involved. The traditional IR model is insufficient for use in studies such as this one.

A number of interactive models for IR have been discussed and proposed (Belkin, Cool, Stein & Thiel, 1995; Belkin & Vickery, 1985; Ingwersen, 1992, 1996), but as yet none is as widely adapted and used as the traditional model. We adopted and adapted a number of concepts from these interactive models to formulate a stratified interaction model of IR for use in our studies. It is briefly summarized below, while it is elaborated in more detail in Saracevic (1996).

We can think of interaction as a series of episodes in time occurring in several connected levels or strata. Each level involves different elements and/or specific processes (a similar notion is used in communication and linguistics under the name of stratificational theory). Here we simplify the stratified models or strata on the user side: surface, cognitive, situational, and affective. The computer or system side has levels as well: surface, engineering, processing, and content.

On the *surface level,* interaction is a series of episodes in time in which users interact through an interface with a system to do not only the searching and matching (as depicted in the traditional IR model) but also engage in a number of other processes or ''things'', above and beyond searching and matching, such as exploring the attributes of a given system, or information resource, navigating, visualizing results, engaging in various types of feedback, and so on; and systems interact with users with given processes and ''understandings'' of their own, and provide given responses in an episode.

Our analysis of search term selection is on such a surface level. In other words, in this study, we concentrate only on the surface level of interaction, while acknowledging (but not investigating) the role of other levels described next.

On the *cognitive level,* users interact with the ''texts'' (including images, data, and their representations) in the information resources, considering them as cognitive structures. Users interpret and judge cognitively the texts obtained, and may assimilate them cognitively. On the *situational level,* users interact with the given situation or problem-at-hand which produced the information need and resulting question. The results of the search may be applied to the resolution or partial resolution of a problem. Users judge the texts obtained according to their utility. On the *affective level,* users interact with intentions and motivations, and associated feelings of satisfaction, frustration, urgency, etc.

However, things are not that simple. The situation that was the reason for interaction to start with produced a problem that sometimes may be ill defined, and the related question, if not on paper then in user's mind, also may be defined in various well-ill degrees. A user also brings a given knowledge state, belief, intent, and motivation related to the situation. Trying to capture all these is called *user modeling,* a diagnostic process that has not been mastered well in automated IR or any other computer application, such as AI. All this is used on the surface level to select files, search terms, search tactics, and other attributes in searching and decision-making and on the deeper, cognitive level to interpret and otherwise cognitively process the texts, and make relevance judgments and other decisions.

During IR interaction, as the discourse progresses through episodes, these deeper level cognitive, situational, and affective aspects in interaction can and, often do, change—the problem or question is redefined, refocused, satisfaction or frustration sets in. Thus, as the interaction progresses, things on the surface level change as well: e.g., new search terms are selected, old abandoned, tactics are adapted and changed, and so on. There is a direct interplay between deeper and surface levels of interaction. Search term selection from different sources and at different episodes (stages of the process, e.g., pre-on-line and during on-line episodes), reflects such interplay. The interplays explain changes in search term selection. Understanding interaction requires understanding these interplays.

The intervention of an intermediary in the interaction process (such as in mediated on-line retrieval) adds still another complex stratum, very interesting in itself. The roles that intermediaries play can also be decomposed into levels. On the surface level, intermediaries use their mastery (knowledge and competence) of IR systems—their contents, techniques, peccadilloes—not mastered by users. This is used to provide effective interaction with the system on the surface level. But on the deeper or cognitive level, they also provide clarifying and diagnostic aspects. They provide help in defining the problem,

TABLE 1. Summary of the data corpus.

Questions	
Number of questions (1 per user)	40
Hours of videotapes	46.05 hours
Mean time per question	69.08 min
Mean time: presearch interview	13.04 min
Mean time per on-line search	56.04 min
Search Intermediaries	
No. of search intermediaries	4
Mean experience per intermediary	8.5 years
Items Retrieved	
Total no. of items retrieved	6225
No. of relevant (R) and partially relevant (PR) items retrieved	3565
No. of not relevant items retrieved	2660
Mean precision per question	57%
Min/Max of total items retrieved	13/427
Standard deviation	85.9
Min/Max of R + PR items retrieved	1/348
Standard deviation	71.43
Min/Max of not relevant items retrieved	0/180
Standard deviation	47.2
Databases Searched	
No. of different databases searched	46
Mean number of databases searched per question	3

focusing the question, and incorporating the context as well as other aspects that enter into user modeling. During subsequent episodes, as the interaction and search progresses, intermediaries may also suggest changes in the problem or question definition. All this plays a critical role in the selection of search aspects on the surface level: files, tactics, and search terms in particular. Thus we treat the selection of search terms as one of the dynamic interactive processes in IR. The selection process is realized and manifested on the surface level, while the effectiveness of search terms, involving user relevance judgments, is established at the cognitive and possibly also situational levels.

Methodology for Data Collection and the Data Corpus

Data used in this analysis were collected during a larger investigation of mediated on-line searching, reported elsewhere (Saracevic et al., 1989, 1991). Because these papers contain a fuller description of the project as a whole, we will only briefly summarize the project and data involved.

Table 1 provides an overview of the data corpus from the larger study. Forty self-selected academic users (faculty and doctoral students) with real information problems provided one question each for on-line searching on DIALOG. Four professional search intermediaries were involved, each doing 10 questions. The 40 questions included topics in medicine, social sciences, physical sciences, and the humanities, and searching as many DIALOG databases as necessary for a given question. Users filled out a standardized form for the question statement before the search, listing the title and description of their question, as is customary in the procedures of the Rutgers University libraries.

The interaction between users and intermediaries was videotaped during a pre-on-line search interview and during the actual on-line search. The transaction logs of the searches were recorded. The discourse between the users and intermediaries was transcribed from the videos. Later, utterances in the discourse from the transcripts and entries (commands, responses) from the transaction logs were synchronized as to time of appearance in the total interaction to allow for various analyses. The written user questions and the synchronized transcripts and logs were the basic data for determining search term sources and for the analysis of sequences. The transaction logs also provided data for the determination of *cycles* (set of commands between the beginning of a search and the display (printing, viewing) of retrieved items, or between two display commands), and moves defined above by Fidel (1991) study. The variables, moves, and cycles are mentioned later in the paper when reporting on the correlation with search term sources as variables.

Users were provided with a printout of all items retrieved during and/or at the end of the search, consisting of the full item, as a database provided. Each user examined each retrieved item and judged it as being relevant (R), partially relevant (PR), or not relevant (NR). We provided users with an instruction sheet defining the meaning of relevance as being topical relevance; however, we can not tell whether users used their own and/or other criteria for relevance judgments. These relevance judgments were used to determine the effectiveness of various search term sources. In addition, users completed an extensive questionnaire for each question involving, among others, a score on a Likert scale related to a user's broad conceptual knowledge of the domain; specific knowledge or expertise on the problem-at-hand; familiarity with the language or terminology used in the problem or domain; and overall satisfaction with the results of the search.

These variables are also correlated with search term sources. In this, and all other analyses, we collapsed the users relevance judgments from three to two classes: items judged relevant (R) and partially relevant (PR) were collapsed into one category (R + PR), from now on simply called *relevant.* In other words, we treated relevance as ''weak'' not ''strong'' relevance (Saracevic & Kantor, 1988). This was done for two reasons: expediency in calculation and, even more importantly, isolation of whatever may be relevant to any degree from that which was definitely judged not relevant.

Methods of Analysis

Identification of Search Term Sources

For the analysis of sources of search terms and their effectiveness, we used the following data: users' question statements; transcripts of the user-intermediary discourse; transaction logs; and users' relevance judgments. By using these together, we can credit each search term with its source and each *search statement* (a set of search terms connected with a logical operator, AND, OR, NOT) where a given search term appears with the retrieval of a given item judged relevant or not relevant. We can also identify the sequence in use of search terms. After preliminary analysis of all the data, (using grounded theory approach as described by Corbin & Strauss, 1990), we identified all the possible sources of search terms. There were five sources: Question Statement, User Interaction, Thesauri, Intermediaries, and Term Relevance Feedback. They are defined in Figure 1. All later analyses refer to these five sources.

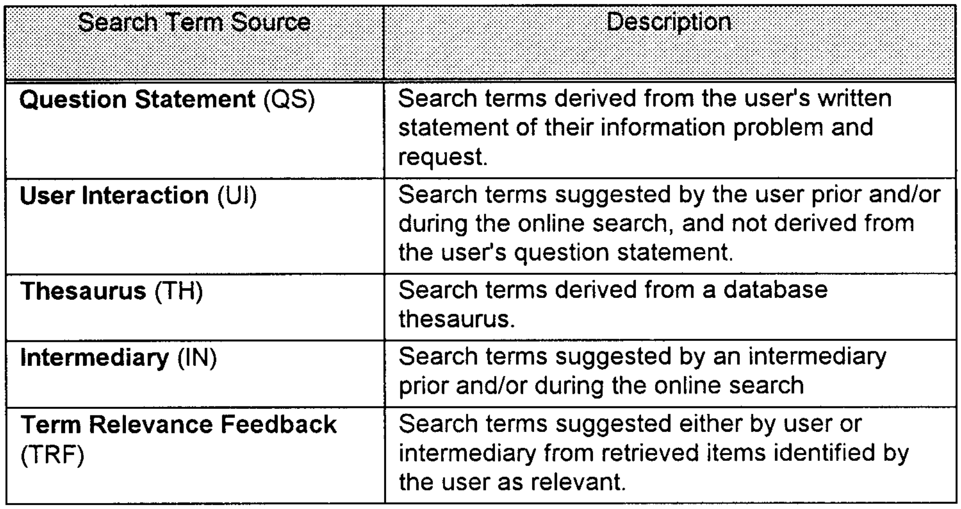
The following method was used in determining and categorizing the sources of search terms. First, for each of the 40 questions, the total set of search terms was identified and listed. Second, using the user's written question, the transcript of the user-intermediary dialogue (or where necessary the videotape itself), and the synchronized search log, the *first mention* of the search term was identified as to its source. Third, the presence of a search term in each item judged relevant (R + PR) and not relevant (NR) was recorded. The search term(s) that were responsible for the retrieval were credited so. In other words, a search term that appears in an item (as given by a database) was credited with the retrieval. If a search term was specified to appear only in a given field, only that field was examined; e.g., if a search term was searched as a descriptor only, then only the descriptor field was examined for that term and not the other fields. Fourth, the number of search terms categorized by sources was summed over the 40 searches.

To provide some clarification, the user-written question statement was submitted at the outset of interaction; thus it is the starting point for crediting it as a source of a search term. As the interaction progressed, other sources of additional search terms were credited in order of appearance. This should be stressed: we credited the source where a term was mentioned first, irregardless if at a later time the term was again and again identified through another source. Thus, for instance, if a search term was first mentioned in the user's question statement, we credited the question statement, and thus the user, irregardless if that term was later also found in a thesaurus or was frequently mentioned by the intermediary. In particular, we did not analyze the use and effectiveness of thesauri and controlled vocabulary searching. We only credited a thesaurus as a source, if, and only if, in the course of events or episodes, a thesaurus was chosen for elaboration on what to search on and as a result a new search term was selected that was previously not selected by another source. (If we used another policy, e.g., where the user is not required to submit a written question, but s/he presents it verbally during interaction with an intermediary, our results may or may not differ—we don't know). Data on these aspects are presented in Tables 2–4.

Effectiveness of Search Term Sources in Relation to Relevance

As mentioned, a major objective was to examine the relationship between the search terms and their sources on the one hand and retrieved items as judged relevant or not relevant by users on the other hand. In this analysis, we took a two-pronged approach. In the first approach, we concentrated on the sources of search terms that produced relevant items only, as presented in Tables 5 and 6. For brevity, we refer to terms from a given source that were responsible for retrieval of relevant items as *relevant terms.*

In the second approach, we analyzed not only search terms that were responsible for the retrieval of relevant items but also those that were responsible for the retrieval of not relevant items, or had no retrieval. We classified the search terms for a given question as to the retrieval of items judged for relevance in four effectiveness categories as defined in Figure 2: terms that retrieved relevant items only; terms that retrieved both relevant and not



Search Term Source	Description
Question Statement (QS)	Search terms derived from the user's written statement of their information problem and request.
User Interaction (UI)	Search terms suggested by the user prior and/or during the online search, and not derived from the user's question statement.
Thesaurus (TH)	Search terms derived from a database thesaurus.
Intermediary (IN)	Search terms suggested by an intermediary prior and/or during the online search
Term Relevance Feedback (TRF)	Search terms suggested either by user or intermediary from retrieved items identified by the user as relevant.

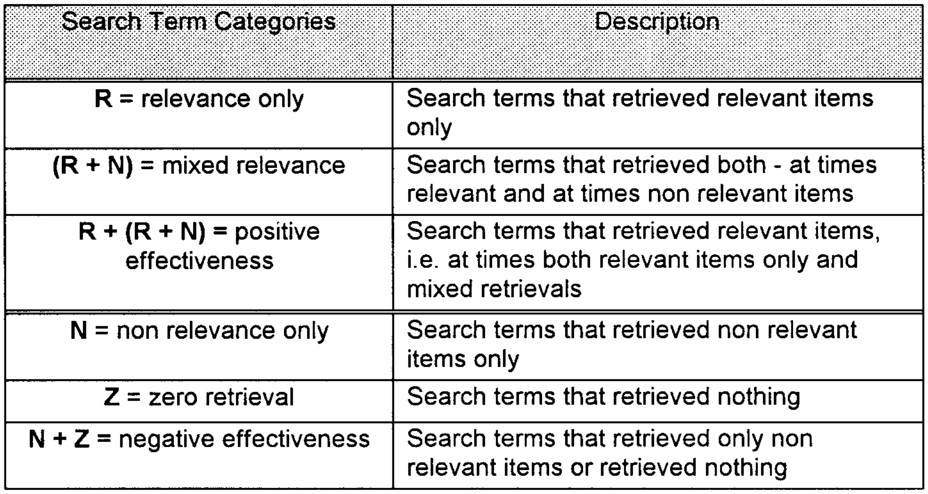
FIG. 1. Definition of sources of search terms.

relevant items; terms that retrieved not relevant items only; and terms that had zero retrievals (e.g., some terms in a logical OR sequence can have zero retrievals). The second category is a recognition that there are search terms that are responsible for retrieval of both items judged relevant and at the same time also for other items judged not relevant. To get a more general picture of effectiveness, we collapse the effectiveness categories from four to two: *positive effectiveness* by adding terms in categories R + (R + N), i.e., we combine terms that produced any relevant items; and *negative effectiveness* by adding terms in categories N + Z, i.e., combining terms that retrieve only not relevant items or had no retrievals at all. In this way, we can separate clearly those terms that produced any relevant items from those that were total duds. The results are presented in Tables 7 and 8.

Proportional Term Weighting

The preceding procedure for identifying and categorizing sources did not include the consideration of the logical connectors (AND, OR, NOT) between terms in search statements with more than one search term. Thus, a proportional weighting scheme for search terms in a statement was developed. The idea was to normalize by weighting the retrieval contribution of all relevant and not relevant items, taking into account the logical connections between terms in a search statement, and then to add the total weight for all search terms in a statement retrieving a given item—be it relevant or not relevant. We started by taking the position that the total weight for all search terms in a search statement in relation to retrieval of an item equals one. This weight then is distributed among search terms based on the given logical combination connecting the terms. In other words, each item retrieved (relevant or not relevant) by a search statement had a total weight of 1; each term contributing to the retrieval of that item received a proportion of 1, according to the following rules:

(1) If only one term in a search statement is responsible for the retrieval of an item, its retrieval weight is 1.



Search Term Categories	Description
R = relevance only	Search terms that retrieved relevant items only
(R + N) = mixed relevance	Search terms that retrieved both - at times relevant and at times non relevant items
R + (R + N) = positive effectiveness	Search terms that retrieved relevant items, i.e. at times both relevant items only and mixed retrievals
N = non relevance only	Search terms that retrieved non relevant items only
Z = zero retrieval	Search terms that retrieved nothing
N + Z = negative effectiveness	Search terms that retrieved only non relevant items or retrieved nothing

FIG. 2. Categories of search term sources according to the effectiveness of retrieval.

(2) If more than one term is responsible for the retrieval of an item, each term received a proportional retrieval weighting. If the search statement connects search terms with a logical AND, (e.g., apple and orange), both terms are considered to contribute equally to the relevant retrieval and both receive a proportional retrieval weighting of 0.5. (Adjacency command between terms was treated as AND). If three search terms contributed to the retrieval of the item, each search term received a proportional retrieval weighting of 0.33.

(3) If two terms are linked by logical OR (e.g., apple or orange) and only one term appeared in the retrieved item, that term was weighted 1 (if apple appeared in the item and orange did not, then apple received weight of 1 and orange received nothing). If both terms appeared in the item, each term received 0.5 retrieval weighting, as neither term could be excluded as having responsibility for the retrieved item.

(4) Search terms excluded by the logical NOT e.g., (apple and orange not pear), are considered to contribute to retrieval and thus, they also receive a proportional weighting, e.g., pear is weighted 0.33. The case of NOT gives special problems to any weighting scheme. We took the position that pear contributed to the retrieval by its very absence in an item, thus in this weighting scheme, it was credited even while not present in the retrieved item.

For each search term, the proportional weights are summed over the number of items retrieved. So if apples retrieved 5 items with proportional weights of 1, 0.33, 0.5, 0.33, and 1, then the proportional weight for apple for the whole search is 3.16. Note that the source of apple and other terms was already determined, thus we can determine the proportional weight numbers and relative percentages of the sources for all the terms in the search and then sum them for the 40 questions. The results are presented in Table 9.

Transition in Sequence of Search Term Use

The objective of this analysis was to determine if regularities versus randomness existed in the transition from the use of one search term (categorized by a given source) to another search term (from whatever source), in the sequence of search terms as they were used during the search process. To observe the transitions, we applied the log-linear analysis, using methods described by Knoke and Burke (1980), and specific refinements and interpretations developed by Mokros (1984). Log-linear analysis is a powerful method to examine the patterns and dynamics of sequences in a process. In this study, log-linear analysis was used to examine the patterns of change between the use of search terms as to their sources and determine whether significant patterns exist in the sequence of terms used at time T and those used at time T + 1. The results are presented in Tables 10 and 11.

Relationship between Search Term Variables and Other Variables

The project as a whole involved the collection of data on a number of variables that characterize the users, questions, interaction, searches, and search outcomes. In this paper, we report on the variables related to search terminology. However, we performed correlation analysis (Williams, 1992) of the variables related to search terminology with many of the other variables in the project, with particular concentration on the user satisfaction ratings; search outcome variables: number of relevant and not relevant items retrieved and precision; search process variables: cycles and moves, as defined above; and user characteristics, such as domain knowledge, also as defined above. For example, we asked questions such as: *Did questions using high number of Question Statement terms result in high precision?*

The specific aim of this analysis was to identify those search term variables that are significantly related to a number of other variables to serve as hypotheses for further study. In this paper, we report on the significant correlation only, omitting those that were not significant. The results are presented in Table 12, where we present only the variables with statistically significant correlation.

Results

Sources of Search Terms

The total number of search terms used in the 40 questions was 593, with a mean of 15 per question, a maximum of 43, a minimum of 4, and a standard deviation of 8.77. The variation in the number of search terms across questions was relatively large, probably reflecting the variety in the complexity of the questions submitted. The distribution of these 593 search terms across the five sources is presented in Table 2. The basic results are interesting from a number of perspectives. Users were responsible for 61% of terms, which means that 39% came from other sources during the interactive process. In other words, users were *not* responsible for 39% of search terms, which is a large percentage. Actually, the role of interaction was even larger: of the 361 user terms, 227 (or 38% of the total of 593 terms) came from the written Question Statement and 134 (23% of total) came from the User Interaction, that is, they were generated during the interaction. Thus, if we add these 23% to the rest of the terms, *we can attribute 62% of search terms to the interaction processes.* This indeed illustrates in a simple but powerful way the contribution of interaction to search term selection.

Thesauri contributed 19% of the search terms. (As a remainder: we looked for the first appearance of a search term as to its source; if a search term was at any time in the process of being searched as a descriptor from a thesaurus but first was identified by some other source, then the thesaurus was *not* credited. Thus 19% does not repre-

TABLE 2. Number and percentage of total search terms selected from each source.

	Total terms	
Sources	Number	Percent of total terms
Users		
Question Statement (QS)	227	38
User-Interaction (UI)	134	23
Sub-Total Users	361	61
Thesaurus (TH)	113	19
Term Relevance Feedback (TRF)		
Selected by users	25	4
Selected by intermediaries	42	7
Sub-total TRF	67	11
Intermediary (IN)	52	9
TOTAL	593	100

Number of questions, 40.

sent the frequency of thesaurus usage in search statements, it represents the sources of search terms first identified through a thesaurus).

Term Relevance Feedback (TRF) contributed 11% of terms, i.e., when some relevant items were retrieved, their examination suggested 11% of the new, additional terms. Out of the 67 terms selected from this source, 25 (37%) were selected by users and 42 (63%) by intermediaries, thus, intermediaries played a significant role in feedback and the selection of new terms from retrieved items (Spink, 1994, 1995). However, using the Term Relevance Feedback results in a broader context, there was a lot of examination of retrieved items as to their relevance during the interaction. In searches involving the 40 questions, Spink (1997) identified a total of 354 feedback loops where users together with intermediaries examined retrieved items for relevance and an additional 67 loops where the items were examined for search terminology. *Relevance feedback clearly played a role in selection of additional search terms, but with 11% of new search terms, this is not as large a role as may be expected.* The examination of relevance during the interactions clearly served additional roles, not only as suggestions for new search terms. This raises an interesting research question: *What is happening during relevance examination in IR interaction?* (As an aside: we do not know what percentage of the search terms are generated by various automatic feedback techniques mentioned above, but it may be of interest to compare them with our data. *Are the percentages comparable?*)

Finally, intermediaries contributed, on their own, 9% of terms—a relatively small amount (this is independent of attribution to intermediaries of terms selected from Term Relevance Feedback—there we took the stance that the retrieved items were the sources and not who selected them from those items). However, the impression of a ‘‘small amount’’ may be misleading because the primary role of intermediaries was not the selection of search terms but the diagnosis and conduct of an effective search and interaction. As shown in the preceding paragraph, intermediaries also contributed significantly to the selection of search terms during relevance feedback. We do not know what would have happened if there were no intermediaries to affect term selection from any other sources, but clearly they played a significant role in the identification and selection. As yet it is not clear at all to what extent such a role could be played for end users by an automated process, i.e., an ‘‘intelligent’’ interface. At minimum, such an interface would have to guide users in the selection of search terms beyond some 38% of terms that the users originally brought to the process.

Distribution of Sources Across Questions

In Table 3, we present the number and proportion of the total of 40 questions that used given search term sources.

Because a written question statement was a starting point for all 40 searches, it is not surprising that all of them had the Question Statement as one of the sources. However, not all the other sources were used in all questions. User Interaction as a source was found in 70% of questions, Term Relevance Feedback in 55%, and Intermediaries and Thesaurus, each, in 50% of questions. Users were the largest contributors in generating search terms, as found above: in close to two-thirds of the questions they also contributed terms during the interaction. Looking at this other way, half of the questions did not use Intermediaries or Thesaurus as sources, and 45% did not use Term Relevance Feedback. These are fairly high percentages in *not* using the specific sources. Most interestingly, *relevance feedback generated new search terms in slightly more than half of the questions; as a rule it was not a regular contributor of search terms.*

However, the picture is a bit more complex. The use of a combination of sources across questions is illuminated in Table 4. The table shows how many sources in a variety of combinations have been used across the 40 questions.

Only 7% of questions used Question Statement alone. This again shows that in an overwhelming majority of

TABLE 3. Distribution of sources across question: Number of questions in which each search term source was used.

	Used in questions	
Search term source	Number	Percent of all questions
Question Statement	40	100
User-Interaction	28	70
Term Relevance Feedback	22	55
Intermediary	20	50
Thesaurus	20	50

Number of questions, 40.

TABLE 4. Distribution of sources across questions: Combinations of sources used to select search terms in given number of questions.

Sources	Used in no. of questions	Percent of all questions	Cumulative, %
Single sources of search terms			
QS alone	3	7	7
Two sources of search terms			
QS/UI	3	7	
UI/TRF	2	5	
UI/TH	1	3	
UI/IN	1	3	
Subtotal	7	18	25
Three sources of search terms			
UI/QS/TRF	4	10	
UI/QS/IN	4	10	
UI/QS/TH	3	7	
QS/IN/TRF	1	3	
QS/TH/TRF	1	3	
Subtotal	13	33	58
Four sources of search terms			
UI/QS/TRF/TH	3	7	
UI/QS/TH/IN	3	7	
UI/QS/IN/TRF	2	3	
QS/TH/TRF/IN	1	3	
Subtotal	9	22	80
Five sources of search terms			
UI/QS/TH/TRF/IN	8	20	100
Total	40	100	

Number of questions, 40; QS, Question Statement; UI, User Interaction; IN, Intermediary; TH, Thesaurus; TRF, Term Relevance Feedback.

questions interaction played a role in the selection of other search terms. In 18% of questions, there were two sources used; in 33% three sources; in 22% four sources, and in 20% all five sources. *It is interesting to observe that all five sources were used in only one-fifth of the questions.* Thus, the variety of sources used in various questions is quite variable. In all probability, this is dependent on the characteristics of the question, such as broad or specific, but we have not investigated this aspect of the data.

Effectiveness of Search Term Sources in Retrieval of Relevant Items

In this analysis, we concentrate on the relationship between search term sources to items judged relevant by users. (As mentioned, we call terms that retrieved relevant items *relevant terms.*) Table 5 shows how the five different sources contributed to retrieval of relevant items: the first column lists the sources; the second gives the number of relevant terms from each source and in parentheses the total number of terms for the source (as given in Table 2); the third column has the percentages of relevant terms from each source (i.e., percent of 378 relevant terms); and the last column provides the percentage of relevant terms for each source in relation to total number of terms for the source (e.g., for Question Statement 81% out of the total number of 227 terms produced relevant answers).

Of the total of 593 search terms, 378, or 64%, of terms retrieved relevant items with a mean of 9 per question, a maximum of 25 and a minimum of 1. The rest of the terms, or 36%, did *not* contribute to retrieval of relevant items, as will be further elaborated in the next section. Search terms selected from users' Question Statements were the most productive in the retrieval of relevant items; while they constituted 38% of the total search terms (see Table 2), they were responsible for close to half (49%) of terms retrieving relevant items. User Interaction generated another 19% of the relevant terms; thus users were responsible altogether for 68% of relevant terms. Term Relevance Feedback contributed 47 or 12% of 378 relevant terms; of the 47 terms, users selected 15 (32% of 47) and intermediaries selected 32 (68%); thus intermediaries were better at spotting potentially relevant terms. Thesaurus was responsible for 14% of relevant terms, while Intermediaries contributed 6% of such terms.

Let us now concentrate on the last column of Table 5: it shows the percentage of relevant terms from each source (i.e., percent of the total number of terms for the source that contributed to the retrieval of relevant items). This analysis allows a further comparison of the effectiveness of search terms from different sources. Most Question Statement terms (81%) contributed to the retrieval of relevant items—this is more than any other source. In comparison, 71% of Term Relevance Feedback terms, 52% of User Interaction terms, 46% of Thesaurus terms,

TABLE 5. Effectiveness: Number of search terms from each source retrieving relevant items (relevant terms).

	Relevant terms	
Sources	Number of relevant terms (total no. of terms for a source)	Percent of total relevance terms (percent of relevant terms for a source)
Users		
Question Statement (QS)	185 (227)	49% (81%)
User-Interaction (UI)	70 (134)	19% (52%)
Subtotal Users	255 (361)	68% (71%)
Thesaurus (TH)	52 (113)	14% (46%)
Term Relevance Feedback (TRF)		
Selected by users	15 (25)	4% (60%)
Selected by intermediaries	32 (42)	8% (76%)
Subtotal TRF	47 (67)	12% (70%)
Intermediary (IN)	24 (52)	6% (46%)
Total Terms	378 (593)	100% (64%)

Total number of search terms, 593.

and 46% of Intermediary terms were relevant terms. Thus, the proportion of relevant terms differ significantly from source to source.

The preceding table deals with the number of search terms from different sources responsible for retrieval of relevant items. In Table 6, we expand the analysis to incorporate the actual number of relevant items retrieved by different sources used singly or in combination in search statements. To simplify, in this table only, we combined sources Question Statement and User Interaction into one category called User. For example, the total number of relevant items was 3,565 (see Table 1); search statements that have User as a sole search term source retrieved 1,801 of these relevant items, while search statements that have User and Thesaurus as a source retrieved 911 relevant items.

Half of the 3,565 relevant items were retrieved by search statements with terms from only one source: the User. However, the other half of the relevant items came from search statements consisting of other sources alone or in combination with each other, including User. This is another way to illustrate the power of interaction and the power of sources of search terms other than User. *Half of the relevant items were retrieved by search statements incorporating user-generated search terms alone, but another half were retrieved by search statements incorporating search terms from other sources in various combinations with user terms or terms from other sources.*

Interestingly, 25% (half of that other half) of relevant items were retrieved by search statements that combined User and Thesaurus as a source, but only 7% that combined User and Term Relevance Feedback. Term Relevance feedback alone or in any combination with other sources in search statements produced only 531 or 15% of the relevant items (recall from Table 2 that Term Relevance feedback accounted as a source for 67 or 11% of all terms). Thus, no matter how we look at it relevance feedback, while being a factor, proportionally it was not a large factor in retrievals. Surprisingly, it was not used much and it did not produce much. The rest of sources and their combinations in search statements produced smaller proportions of relevant answers.

Search Term Sources across Effectiveness Categories

In the preceding section, we dealt with the retrieval of relevant items only. Now we are expanding the analysis to deal with the retrieval of any kind of retrieved items according to effectiveness categories, as defined in Figure

TABLE 6. Effectiveness: Number of relevant items retrieved by each source and combination of sources in search statements.

	Relevant items retrieved	
Sources	Number	Percent of total of relevant items
Single sources		
User (QS + UI)	1801	50%
TRF	39	1%
TH	39	1%
IN	0	0%
Subtotal single sources	1879	53%
Two sources		
User/TH	911	26%
User/TRF	267	7%
User/IN	202	6%
TH/TRF	35	1%
IN/TH	16	0.4%
Subtotal two sources	1431	40.4%
Three or more sources		
User/TH/TRF	148	4%
User/IN/TH	65	2%
User/IN/TRF	13	0.4%
User/IN/TH/TRF	29	0.8%
Subtotal three + sources	255	7.2%
Total Relevant items	3565	100%

Total retrieved items, 6225; Items judged relevant by users, 3565; QS, Question Statement; UI, User Interaction; IN, Intermediary; TH, Thesaurus; TRF, Term Relevance Feedback; User, QS + UI.

TABLE 7. Effectiveness: Number of search terms in each effectiveness category.

	Search terms	
Effectiveness category	*n*	Percent
Relevant items only (R)	25	4
Relevant & Non relevant items (R + N)	353	60
Subtotal R + (R + N) Positive Effectiveness	378	64
Nonrelevant items only (N)	147	25
Zero retrievals (Z)	68	11
Subtotal N + Z Negative Effectiveness	215	36
Total	593	100

2. Table 7 shows the number of search terms associated with each of the four effectiveness categories, e.g., it shows that there were 25 terms that were responsible for the retrieval of relevant items only and 68 terms that did not retrieve by themselves any items—some other term or terms in the search statement where these zero terms were used were responsible for the retrieval.

Even this basic data reveals some interesting aspects of retrieval effectiveness. Note that a very small percentage, 4%, of the total of 593 search terms produced relevant items only. By far the largest percent of search terms, 60%, was responsible for the retrieval of both relevant and not relevant items. *In other words, for well over half of the terms, relevance goes both ways: some items retrieved by the same terms are relevant, others are not.* Of course, this presents a major problem in professional practice and in retrieval algorithms in the selection of search terms, because while the judicious selection of search terms greatly affects retrieval of relevant items, there are other aspects in the items themselves that are important in relevance judgments. It is trivial but still important to observe that yes, without given search term(s) a relevant item cannot be retrieved, however, the mere presence of a search term is not the only variable in determining relevance. The nature or content of retrieved items is also a significant variable and so are a number of other variables. This may be disappointing, but no matter how judicious, the selection of search terms alone does not determine relevance—the majority of the same terms will retrieve both relevant and not relevant items. But judicious selection may sharpen to some extent the precision and recall. Altogether, some 64% of search terms was associated with positive effectiveness.

A significant 25% of search terms produced not relevant answers only, while 11% of terms retrieved nothing at all. Thus, 36%, or more than one-third, of terms have been associated with negative effectiveness. Thus, one-third of terms retrieved only not relevant items or failed to produce any retrievals at all! *For us, this high proportion of negative effectiveness was a surprise; we consider this one of the most significant findings in the study.* This finding supports the hypothesis, long held, that interactive IR is to a significant extent a trial and error procedure (Swanson, 1977).

These results are further amplified in Table 8 where the sources of search terms are associated with the four effectiveness categories. The table shows, for instance, that Question Statements contributed a total of 227 search terms; of these, 11 (5%) were responsible for retrieval of relevant items only and 174 (77%) for retrieval of both relevant and not relevant; thus 185 (82%) of Question Statement terms were associated with positive retrieval, while 42 or 18% were responsible for negative retrievals. Obviously, the last column in Table 8 provides the same data as in Table 2, and the last row the same data as in Table 7.

Each source produced terms that fell within each effectiveness category, but sources differed in their contribution. A chi-square analysis of Table 8 (excluding columns for positive and negative retrieval) proved significant at $P < 0.0000$ with 12 degrees of freedom, showing significant differences in the distribution of search terms and their sources within the four effectiveness categories.

Let us now concentrate on positive and negative effectiveness. Considering the total terms within each category, Questions Statement terms had by far the highest positive (82%) and the lowest negative effectiveness (18%). About half of User Interaction terms had positive and the other half negative effectiveness. *Thus, the users were most effective in term selection when they had to write the question down, suggesting that this is a good practice in general regardless if intermediaries are involved.* Relevance Feedback terms also had a high (70%) percentage of its terms producing positive effectiveness—when selected (and not many were) Relevance Feedback Terms are effective in a positive way more than two-thirds of the time. In contrast, sources Intermediary and Thesaurus had each less than half (46%) of terms associated with positive retrievals. Thus there is a considerable difference between sources in positive and negative effectiveness. However, we cannot assume that all negative effectiveness terms are ''bad'' in themselves. Trial and error with those often leads to selection of other terms that have positive effectiveness.

Proportional Term Weighting in Retrieval

This analysis took into account retrieval of all items (relevant and not relevant) by the search statements as a whole. That is, it took into account the logical connections between terms by proportionally distributing a weight among terms in a search statement responsible for retrieval of all items according to rules presented in the section *Methods of Analysis.* Table 9 contains an analysis of the proportion (percentage) of the contribution of terms from each source in search statements for each of the 40 questions. We provided here, for the first time, detailed analysis, question by question, to illustrate the mentioned wide variation among questions, in addition to providing a mean in the last row. The table shows

TABLE 8. Effectiveness: Number of search items from each source in each effectiveness category.

	Effectiveness categories						
Term source	R (%)	R + N (%)	Subtotal positive effectiveness (%)	N (%)	Z (%)	Subtotal negative effectiveness (%)	Total (%)
QS	11 (5)	174 (77)	185 (82)	30 (13)	12 (5)	42 (18)	227
UI	7 (5)	63 (47)	70 (52)	51 (38)	13 (10)	64 (48)	134
IN	3 (6)	21 (40)	24 (46)	16 (31)	12 (23)	28 (54)	52
TH	2 (2)	50 (44)	52 (46)	34 (30)	27 (24)	61 (54)	113
TRF	2 (3)	45 (67)	47 (70)	16 (24)	4 (6)	20 (30)	67
Total	25 (4)	353 (60)	378 (64)	147 (25)	68 (11)	215 (36)	593 (100)

Positive effectiveness, R + (R + N). Negative effectiveness, N + Z. Percentages of positive and negative effectiveness (in parentheses) relate to total number of terms for a source. Total number of search terms, 593.

for instance, that in Question 2, 100% of proportionally weighted terms came from Question Statement as a source, and in Question 18, 52% came from Question Statement, 31% from User Interaction, and 17% from Intermediary. (Note that there are 40 questions used in analysis, but they are not sequentially numbered).

To summarize the data: Question Statement weighted terms that were responsible for retrieval appear in 37 (93% out of 40) questions; User Interaction in 26 (65%), Thesaurus and Intermediary each in 19 (47%), and Term Relevance Feedback in 14 (35%) of questions. The rank and relations coincide with those reported in Table 3 where all sources of search terms were considered, regardless of weighting.

Let us now concentrate on the means (the last row). In weighted proportion the Question Statement contributed 70% of the terms responsible for the retrieval of all items. Recalling from Table 8 and the accompanying discussion, 525 out of 593 search terms produced any retrieval (additional 68 terms had zero retrieval which is not counted here); 215 out of these 525 terms or 41% came from Question Statement—this is in straight number of terms without weighting. But when weighting is imposed to search statements, then the average contribution to all retrievals of Question Statement terms jumped to 70%. To provide for comparison for all the sources between the two indicators, the weighted average is given here in the first number, taken from Table 9, and in the parenthesis is the percentage of straight number of terms (without weighting) responsible for all retrievals (i.e., not counting zero retrievals), culled from Table 8: Question Statement 70% (41%), User Interaction 22% (23%), Intermediary 21% (8%), Thesaurus 14% (16%), and Term Relevance Feedback 9% (12%).

Thus, weighting of terms, which takes into account logical connectors and the number of terms responsible for retrieval in a search statement, makes a significant difference for some sources, most notably Question Statement and Intermediary, and much less of a difference for other sources. The Question Statement terms and the terms generated by the intermediaries jumped higher. In particular, intermediaries came up with relatively few terms, but those they came up with, coupled in search statements, produced about one-fifth of all retrievals. This illustrates another role of intermediaries that may be very hard to automate.

Order and Transition in Use of Search Term Sources

What was the sequence of search term sources used on a question to question basis? The answer is provided in Table 10. The table lists by columns: question number; number of search terms in each question; number of term sources that provided positive retrievals, i.e., relevant answers; and in the last column the sequence of use of sources of search terms—sources that produced positive retrievals are typed in bold, and sources that produced negative retrievals (not relevant and zero output) are typed in normal font. (Note that data for the total number of items retrieved is presented in Table 2 and for positive and negative retrievals in Table 8).

Table 10 provides details on the use of sources and their effectiveness for each question separately, illustrating again the considerable differences that exist among questions. The lowest proportion of relevant terms in relation to total number of terms was 17% (Question 30), while in eight questions, 100% of terms were relevant ones. Clearly, not all questions are created equal. A research question immediately comes to mind: *what variables are responsible for such individual differences among questions?* Some of their characteristics, such as high or low specificity? Or characteristics of the underlying problem, such as well versus ill defined? We did not address such topics in this study (our objectives were to observe differences, if any, rather than to explain), however, we did illustrate the magnitude of individual differences among questions, which clearly invites further research. The cause of these differences is a significant research question in its own right.

As explained, a written question was submitted by users at the outset. Thus, it is not surprising that the largest proportion of questions, 78% of them, started with a term from Question Statement. This still leaves 22% of the questions that started the searching with a term from some

TABLE 9. Sources of proportionally weighted terms in search statements according to logical connectors and retrieval of items.

Question number	QS, %	UI, %	IN, %	TH, %	TRF, %
2	100	0	0	0	0
3	0	59	2	0	39
4	56	0	0	44	0
5	100	0	0	0	0
6	57	20	11	2	10
7	60	0	0	40	0
8	86	2	12	0	0
9	33	14	14	13	26
10	81	9	0	0	10
11	100	0	0	0	0
12	88	0	0	0	12
14	43	0	26	27	4
15	43	28	0	24	5
16	95	4	1	0	0
17	100	0	0	0	0
18	52	31	17	0	0
19	83	0	0	17	0
20	0	91	0	0	9
21	33	13	0	30	24
22	68	0	0	30	2
24	82	9	9	0	0
25	72	28	0	0	0
26	92	8	0	0	0
27	48	29	2	21	0
28	48	35	0	16	0
29	46	40	11	3	0
30	100	0	0	0	0
31	77	0	0	0	23
32	67	3	3	12	15
33	1	34	16	39	10
34	100	0	0	0	0
35	100	0	0	0	0
36	74	26	0	0	0
37	82	15	0	0	3
38	79	3	0	6	12
39	83	17	0	0	0
40	27	43	3	10	17
41	0	18	0	58	24
42	91	3	0	4	2
43	47	8	10	5	30
Mean	70	22	21	14	9

Percentage indicates the porportion of retrieved items by weighted terms from a given source in a question.

other source. Of 9 questions that did not start with a Question Statement term, 5 (56% of 9) started with a Thesaurus term, 3 (33%) with a User Interaction term, and the remaining 1 was an Intermediary term. An interesting issue (that again we did not investigate): *What sets these Thesaurus-starting questions apart?*

The data from Table 10 was used to analyze, i.e., cross-tabulate, the changes in the sequence of term sources in search statements used at time T and time T + 1. *Did sequential use of terms (distinguished as to their source) in search statements form some structural relationship (model) or follow a random pattern?* For the analysis, we used the log-linear method found in many statistical packages, such as SPSS. In general, the aim of log-linear analysis is to identify the structure or relationships in observed variables, to see whether certain combinations of values are more likely or less likely to occur than others. Possibly, a best fitting model for the data can be specified. Log-linear models try to predict the number of cases in cells of a crosstabulation, based on the values of both individual variables and their combinations. A test of significance assumes no association of occurrences of sources at time T and time T+1. We used a so-called independence model where transitions from the same states (e.g., Question Statement to Question Statement) are taken into account.

Results are presented in Table 11. The first number in a cell is the observed value of the sequence T and T + 1. For instance, the table shows that it was observed that a Question Statement term was followed by a Question Statement term on 140 occasions, by an User Interaction term on 32 occasions, by an Intermediary term on 16 occasions, and so on. Below each observed value provided are the expected value in parenthesis (calculated as sum of row times sum of column divided by total observations), the residual value (difference between the observed and expected value), and standardized residual (residual between observed and expected value divided by the standard deviation of all residuals). Standardized residuals roughly more than +2 or less than −2 indicate significant differences in observed transitions from what was expected at $P < 0.05$. Cells with standardized residual between −2 and +2 are interpreted as random and thus not structural. Cells with a value of greater than +2 may be interpreted as a relationship that occurs at a greater than chance rate and a value less than −2 as a relationship that occurs at a rate less than chance. The whole may be thought of as a structure with various relationships between members (or in our case between sources). A value more than +2 can be considered as a facilitating relationship, while a value less than −2 as a prohibiting one.

Log-linear analysis of the sequences of search term use resulted in a likelihood ratio chi-square of 276.615 with 16 degrees of freedom. This finding is significant at the 0.05 level. Thus, we can reject the null hypothesis of no association between sequences. But can we say something more? Certainly:

(1) Sequences that occurred more than expected were: Question Statement (QS) to QS (not surprisingly given their high number to start with); User Interaction (UI) to UI; Intermediary (IN) to IN; Thesaurus (TH) to TH; Term Relevance Feedback (TRF) to TRF. All are diagonal crosstabulation cells. All refer to going from one source to the same source again at a higher then expected rate. For instance, after a Question Statement term is used, it is highly likely that the next term will be again a QS term and less likely that the next term used will be from another source. There is a higher than expected chance that the same source will be followed in the sequence of search terms.

(2) Sequences that occurred less than expected were: QS to UI; QS to TH; QS to TRF, UI to QS; TH to QS;

TABLE 10. Sequential order of search term use classified by source.

Question number	Number of terms	Relevant terms		Order of search term use by source
		n	Percent	
2	4	4	100	**QS QS QS QS**
3	16	9	50	**QS** UI **UI** IN TRF UI UI **UI TRF QS QS** UI **QS** UI **TRF TRF**
4	4	4	100	**TH QS TH QS**
5	10	7	70	QS **QS QS QS QS QS QS** UI UI **UI**
6	15	10	66	QS **QS QS** QS **QS** QS **UI UI TH IN** UI **TRF QS QS** QS
7	13	11	85	**QS QS QS QS QS QS QS TH TH** UI **QS** QS **TH**
8	18	12	67	**QS QS QS IN QS** UI **UI QS QS IN IN QS** QS **QS** TRF UI QS IN
9	14	11	79	**TH TH QS QS** IN **TRF TRF** TH **UI** IN IN **UI IN TRF TRF**
10	19	18	95	**QS QS QS QS QS UI UI UI TRF TRF TRF QS** QS **QS TRF TRF TRF TRF TRF**
11	8	5	63	**QS QS QS** QS QS QS **QS QS**
12	6	6	100	**QS QS QS QS QS TRF**
14	6	6	100	**QS IN TH TRF IN TH**
15	28	5	18	**QS** TH TH TH TH **UI** QS UI UI UI UI UI QS TH UI TH TH UI UI IN **TH** TH TH UI **TRF** QS QS **QS**
16	16	6	38	**QS QS QS QS** QS QS UI UI UI UI UI UI **IN** IN **RF** TRF
17	6	5	83	**QS QS QS QS QS** TH
18	19	14	74	**QS IN UI** UI **UI** UI IN **UI UI UI UI QS IN QS** UI **UI QS** IN **IN**
19	8	6	75	**QS QS QS** IN **QS QS IN** IN
20	28	13	46	UI **UI UI UI** UI UI UI **UI UI** UI UI **UI UI** UI UI **UI** UI **UI** UI UI **UI UI** UI **TRF** UI TRF **TRF** TRF
21	16	15	94	**TH QS TH TH TH UI TRF TRF TRF TH TH TRF QS UI** TRF **QS**
22	7	7	100	**QS QS QS QS QS TRF TH**
24	9	9	100	**QS QS UI QS QS UI QS IN QS**
25	24	16	67	**QS QS QS QS QS** QS QS **UI UI UI UI QS** QS **QS QS QS QS** IN **QS** QS **UI** QS IN IN
26	10	7	70	**QS QS QS** QS IN **QS** QS **QS UI UI**
27	23	21	91	**QS QS QS IN QS IN QS QS UI UI TH TH UI TH** TH IN **TH UI TH QS TH UI TH QS**
28	14	9	64	**UI** TH **TH TH** QS **QS QS UI QS** IN **TH TH** TH TH
29	19	7	37	**QS** QS **QS** QS TH UI **UI** IN **QS TH** TH TH TH TH TH TH TH **IN UI**
30	6	1	17	**QS** QS QS QS QS UI
31	13	9	69	**QS** QS **QS QS** QS **QS** QS QS **TRF QS QS QS QS**
32	30	25	83	**QS QS QS QS QS QS QS QS** QS **QS TH QS** TH **UI** UI UI **QS QS UI UI UI IN TRF TRF TRF IN QS** TH **TRF TRF**
33	23	11	48	**TH** TRF UI **UI** TH TRF TH **TH** UI TH UI UI UI **IN QS UI TH TH** TRF **TH** TRF TRF **IN TH**
34	6	4	67	**QS** TRF TRF **QS QS QS**
35	5	5	100	**QS QS QS QS QS**
36	15	8	53	**QS** UI UI **UI UI QS QS QS QS** TH TH UI QS **QS** QS
37	11	7	55	**QS** IN **QS** QS **UI** QS QS **QS TRF TRF QS**
38	12	10	83	**QS QS QS TRF TRF QS** QS **QS UI** TH **TH TH**
39	5	5	100	**QS UI QS QS UI**
40	19	11	58	**QS QS** UI UI **TRF** UI **UI IN IN UI UI** UI UI UI **TRF TH** UI **UI** TH
41	43	10	23	TH TH TH TH TH TH TH **TH QS** TH **TRF** TH TH UI IN IN TH **TRF TH** IN TH TH UI TH TH **TRF TRF** TH TH TH TH TH **TH** TH **TH** UI **TH** UI TH IN TH TH TRF
42	12	6	50	UI **TH** QS QS UI **QS QS** TH TH **UI QS TRF**
43	27	23	85	**IN IN TH TH TH** TH **TH TH TH** TH TH **QS QS QS TRF QS IN IN UI UI TH UI TRF TRF** TRF **TRF TRF**
Total	593	378		

Boldface items are relevant terms and typed normal are terms that retrieved nonrelevant items or had no retrievals.

and TH to UI. In other words, QS goes back to itself and with less chance to any other source. Once UI is selected, it has less chance to go to QS. TH terms have less chance then to be followed by either QS or UI terms.

(3) The remaining sequences have no structural relationship to each other, they may be considered as random.

Because the diagonal (same to same source) sequences were so overwhelmingly present, we performed another test, using a so called quasi-independence model. In this model, we considered the same to same source transitions as structural zeroes, i.e., we eliminated the values of diagonal cells and recalculated the expected values, residuals and standard residuals. A chi-square test was not significant at 0.05. (Thus, to save space, we are not presenting this table). This means that once we remove the same-to-same source transitions from time T to T + 1, there is no significant association between sources in remaining transitions. As it turned out, the main characteristic of the

TABLE 11. Transition in sequences of search terms by sources.

	Time (T + 1)					
Time (T)	QS	UI	IN	TH	TRF	Total
QS Observed	140	32	16	17	10	215
Expected	(77.7)	(51.2)	(17.5)	(43.2)	(25.4)	
Residual	62.25	−19.17	−1.45	−26.24	−15.39	
St. Residual	7.06	−2.68	−0.35	−3.99	−3.05	
UI Observed	21	72	11	17	13	134
Expected	(48.5)	(31.9)	(10.9)	(26.9)	(15.8)	
Residual	−27.46	40.11	0.12	−9.95	−2.82	
St. Residual	−3.94	7.10	0.04	−1.92	−0.71	
IN Observed	13	8	10	10	4	45
Expected	(16.3)	(10.7)	(3.7)	(9)	(5.3)	
Residual	−3.27	−2.71	6.35	0.95	−1.31	
St. Residual	−.081	−0.83	3.32	0.32	−0.57	
TH Observed	11	22	4	57	11	105
Expected	(34)	(22.4)	(7.6)	(18.9)	(11.1)	
Residual	−22.99	−11.37	−3.63	38.1	−0.1	
St. Residual	−3.94	−2.4	−1.31	8.76	−0.03	
TRF Observed	11	6	3	8	26	54
Expected	(19.5)	(12.9)	(4.4)	(10.9)	(6.4)	
Residual	−8.53	−6.85	−1.38	−2.86	19.62	
St. Residual	−1.93	−1.91	−0.66	−0.87	7.77	
Total	196	140	44	109	64	553

Cell values: value of observed frequency of transition from source at time T to next source at time T + 1; expected value; residual value; and standardized residual. Expected values are in parentheses for differentiation with observed values.

model of sequences of search term sources is that they tend more likely to follow each other than not and that when they are followed by some other source it is more likely to be a random pick.

Relationship between Search Term Variables and Other Variables

As mentioned, the study of search terms reported in this paper was part of a larger study of IR interaction where a number of other variables were observed, above and beyond the search terms. Because we had data on a large number of other variables, we took the opportunity to correlate them with search term variables reported here. That is, we asked the question: *Was there any significant relationship between search term variables and other variables in the study that pertain to search outcomes or user and search characteristics?* We performed Pearson correlation analysis involving 118 pairs of variables, 23 of which were statistically significant. We assumed, as do all Pearson correlations, a linear relation between correlated variables. The significant correlations are presented in Table 12, the nonsignificant relations are ignored. We are fully aware that this kind of blanket correlation among a large number of variables in a study amounts to a ‘‘fishing expedition,’’ frowned upon in statistics. It means that if there are a lot of correlations made, as in this case, some of them may show up as being statistically significant even when there is no relation; a significance level < 0.05 means that we may expect that some 5 out of 100 correlations may show up significant even when they have no relationship. We engaged in the fishing expedition for two reasons: curiosity (could not resist the temptation) and providing the results as hypotheses for further verification and study.

With the caveat that all the conclusions should be treated as hypotheses, we discuss the correlations in six major categories as presented in Table 12.

User satisfaction. Users expressed their satisfaction with the results for their question on a 5-point scale from 1, (very dissatisfied) to 5 (very satisfied). Satisfaction rating was negatively correlated with two search variables: number of Thesaurus terms and number of nonrelevant terms. Thus, questions with a higher proportion of terms selected from a Thesaurus produced lower user satisfaction. While this may be surprising, it does relate to the results in Table 5, showing Thesaurus terms as producing a low percent of relevant terms, and Table 8, which shows that Thesaurus terms had a high percent of negative (nonrelevant and zero) retrievals. It is not surprising that questions with high number of nonrelevant terms related to lower user satisfaction.

Precision. Higher precision was significantly related to greater use of Question Statement terms. As the number of Question Statement terms increased, the precision of the search increased. The terms derived from Question Statements featured heavily in retrieved items judged relevant by users. This supports the findings from Table

TABLE 12. Significant relationships between search term variables and other variables in the study.

Variable	P
User satisfaction	
No. of thesaurus terms per question	−0.0049
No. of nonrelevant terms per question	−0.0029
Precision	
No. of nonrelevant terms per question	−0.0144
No. of question statement terms per question	0.0301
No. of nonrelevant terms	
Percentage of user-interaction terms per question	−0.001
No. of user-interaction terms per question	−0.0279
No. of question statement terms per question	−0.0001
No. of thesaurus terms per question	0.0000
Percent of Question Statement Terms	
No. of TRF terms per question	−0.0000
No. of cycles per question	−0.0006
No. of total terms per question	−0.0000
No. of moves per question	−0.0003
Users broad knowledge domain	0.0275
Users specific knowledge domain	0.0275
User familiar with language	0.0444
No. of TRF Terms	
Percentage of user-interaction terms per question	−0.0010
No. of cycles per question	0.0092
No. of moves per question	0.0034
No. of relevant terms per question	0.0095
No. of Thesaurus Terms	
Percentage of user-interaction terms per question	−0.0000
Percentage of TRF terms per question	0.0389
No. of cycles per question	0.0303
No. of moves per question	0.0001

Only statistically significant correlation at $P < 0.05$ included.

5, showing that Question Statement terms produced the highest proportion of relevant search terms, and Table 8, showing that they had the highest percentage of positive retrievals. Again, it is not surprising that precision does not go along well with a higher number of nonrelevant terms.

Non-relevant terms. This further elaborates on points already made. As the number of nonrelevant search terms increased, the number of User-Interaction and Question Statement terms decreased. Or conversely, when users contributed less, nonrelevant terms rose. An increase in nonrelevant terms was accompanied by an increase in the number of additional terms selected from Thesauri. This finding may suggest that as the terms from users were exhausted and the on-line search continued, other sources, e.g., Thesaurus, were used to identify search terms. Searches with higher precision or user satisfaction used less Thesaurus and nonrelevant terms and more Question Statement terms.

Percentage of Question Statement terms. Questions with a higher percentage of Question Statement terms were also questions with a fewer number of Term Relevance Feedback terms and total number of terms, as well as fewer cycles and moves. This finding suggests that where Question Statement terms predominate, the search process was shorter and to the point and that other sources were less likely to be used to find additional search terms. Conversely, with fewer numbers of Question Statement terms the on-line search was longer and more interactive. Additionally, as the user's broad and specific domain knowledge and language familiarity increased (also scored by users on a 5-point scale), questions included more Question Statement terms. This finding suggests that if a user had a higher domain knowledge and a greater familiarity with the language of the knowledge domain to start with, s/he tended to include more terms that were eventually used in their written question. Such users knew what they wanted from the outset and specified so. In another study, Spink (1993) also found that as the knowledge level of the user increased accompanied by a previous on-line search on the same topic, there was a corresponding increase in precision.

Term Relevance Feedback terms. The number of TRF terms was inversely related to the percentage of User Interaction terms in a question. However, a higher number of TRF terms was also related to a higher number of cycles and moves, i.e., more interactive searches, and a higher number of relevant terms in questions. Not surprisingly, this suggests that TRF terms were selected later in the on-line search process and when there was more interactivity. If the Question Statement and User-Interaction terms were exhausted and the user and intermediary still wanted to continue with the on-line search, the number of moves and cycles and subsequently their use of TRF increased.

Thesaurus terms. Again, this is an elaboration on previous findings. As the number of Thesaurus terms (and nonrelevant terms, as found above) increased, the percentage of User-Interaction terms (and user satisfaction) decreased. However, with an increase of TH terms, the number of TRF terms and the number of moves and cycles also increased. This finding suggests that Thesaurus terms were used with an increase in interactivity, but were also associated with increasingly nonrelevant retrievals.

Conclusions and Implications

In this study, we investigated the sources of search terms involving users and intermediaries in an interactive, on-line information retrieval process. Data were derived from a larger study of IR interaction, designed to observe a variety of interactive aspects and variables in a real-life (as opposed to laboratory) setting. The objectives were to classify the sources of search terms and analyze their behavior and, because we had users' relevance judgments, also to analyze the retrieval effectiveness of various sources. We performed a number of qualitative and statistical analyses on the data. The results of these analyses are presented in the preceding section.

We included a number of pragmatic implications and suggestions for research questions with the results of specific findings. In this section, we suggest a number of more general conclusions and implications. Of course, there is a limit to our conclusions. As mentioned at the outset and again later in the paper, we cannot really claim generalizations beyond our own data and setting, any more than any other case study can. Still we are offering these conclusions to be taken with due caution as possible guidelines in the conduct and instruction of IR processes, as factors to be considered in design of IR interfaces, and, even more so, as hypotheses for further study.

Conclusions on Sources of Search Terms

Question Statement terms. These terms were derived from the written question as submitted by users at the outset of interaction. They formed the largest proportion of total search terms, close to two-fifths of all terms were QS terms. They were used in all 40 searches. The great majority of QS terms (more than four-fifths of QS terms) were also relevant terms, i.e., contributing to retrieval of relevant items; no other source had such a high proportion of relevant terms. In combination with User-Interaction terms, they were responsible for the retrieval of half of the relevant answers.

User-Interaction terms. User-Interaction terms were second in relation to the total number of terms, but, only less than half of UI terms were also relevant terms. As a source, they were used in close to two-thirds of the questions. About half of the UI terms were responsible for the positive and the other half for the negative retrievals.

Thesaurus terms. While they formed about one-fifth of all search terms, they were used in half of the questions. Close to half of the TH terms were also relevant terms, which means that more than a half were associated with negative effectiveness. More often than not, they were used toward the end of a search statement. TH terms proved most effective when combined in search statements with User terms. However, their overall positive effectiveness was lowest, and their negative effectiveness was highest, both spots shared with Intermediary terms. The relative low productivity of thesauri for the suggestion of new search terms and low effectiveness of Thesaurus terms was somewhat of a surprise.

Term Relevance Feedback terms. A little more than one-tenth of terms came from TRF, but they were used in more than half of the questions. Although small in total number of terms and contributing a small percentage of relevant retrievals, when used TRF terms were quite effective in relevant retrievals: close to two-thirds of TRF terms contributed to positive and less than one-third to negative retrievals, second in such percentages to Question Statement terms. As a source of search terms, TRF terms were not used a lot, but when used, they were effective.

Intermediary terms. With less than one-tenth of the total number of search terms intermediaries were the smallest contributors, moreover, of the terms suggested and less than half were relevant. Their terms were used in half of the searches. The effectiveness of IN terms was the same as Thesaurus terms. However, intermediaries played a significant role in the selection of Term Relevance Feedback terms: of the total of TRF terms, about two-thirds were suggested by intermediaries, the rest by users. Of the TRF terms picked by intermediaries, two-thirds were also relevant terms. The role played by the intermediaries was clearly not to generate search terms, but to guide the term selection process and the search interactions as a whole.

Effectiveness of Search Term Sources

Here we look at the same results from the focus of the contribution to positive and negative effectiveness, where, as explained, positive effectiveness means contribution to relevant retrievals, and negative effectiveness means nonrelevant or zero retrievals only. QS terms had by far the highest positive (over four-fifths of the number of QS terms had positive effectiveness) and lowest negative effectiveness, followed by TRF terms, at close to two-thirds of them with positive effectiveness. On the other end were Thesaurus and Intermediary terms with less than half of their respective numbers having positive effectiveness; that is, more than half had negative effectiveness in retrievals.

This brings up an important point. While users contributed most of the search terms and most of the productive terms, in terms of positive effectiveness, they did *not* contribute them all. Other sources were significant for the selection of search terms retrieving relevant items. The finding suggest that users may not have generated the additional non-user search terms on their own or conduct the search alone in its full interactive complexity.

One of the surprising findings was the relative large proportion of terms producing negative effectiveness: more than one-third of all terms produced nothing but nonrelevant answers or had no retrievals at all. Simply put: more than one-third of terms were duds.

Positive effectiveness includes two categories of terms: those that produced relevant answers only and those that produced both relevant and not relevant answers, or mixed retrievals as to relevance. A very small number, less than one-twentieth of all terms, retrieved relevant answers only. But close to two-thirds of all terms retrieved at times relevant and at other times nonrelevant answers. This illustrated the conundrum of IR: *in fairly large numbers search terms can and do go both ways as far as relevance is concerned.* While judicious selection of terms, be it algorithmic or by professionals and knowledgeable users, can improve retrieval effectiveness, there may exist a ‘‘natural’’ limit. Probably, there will be always terms that will go both ways as to relevance. This is because, among others, interaction on the cognitive and situational levels with the nature and content of retrieved items plays a major role. The challenge IR systems designers face is to facilitate the reduction of nonrelevant items in mixed retrievals, and of nonrelevant retrievals, and nonretrievals. However, the reduction of mixed retrievals through automatic processes is not clear. Researchers should consider the extent to which the mixed retrievals occur and how they can be limited through algorithms and user training.

Correlations

We performed a number of correlations of search term variables with variables related to outputs and user characteristics. Although we are cautious with our conclusions, the correlation results confirm other findings.

Thesaurus terms present an interesting and even surprising case: users were less satisfied with higher use of TH terms; the number of nonrelevant terms rose with higher use of TH terms, and so did the percentage of Term Relevance Feedback terms and number of moves and cycles, while the number of User Interaction terms fell. This should *not* be construed as indictment of thesauri. Clearly, the more interactivity in a question, the more use of Thesaurus terms. But an increase in interactivity may also mean a less focused question to start with and more necessity to probe for something that was not clear from the outset. In such a case, a thesaurus may be a last resort, where nothing may help to produce more satisfactory results. Thesauri are extremely important tools in IR. This study illustrated how they are used and with what effects, but the use of thesauri in searching warrants further research, particularly in relation to making them a more readily and more positively used tool for search term selection.

Correlations also further illustrated the significance of Question Statement terms. When users knew a lot about the domain and language of the question, they constructed a Question Statement with specific terms that later proved to be pivotal in the selection and effectiveness of search terms. With an increase in number of QS terms, came an increase in precision and a decrease in the number of nonrelevant terms. This brings up a larger point and implication. In our model of interaction, users were required to bring a written question. The language of the written question proved to be highly significant for the selection of search terms and their effectiveness. Thus, it may be advisable to suggest to users involved in either mediated or nonmediated searches to prepare a written question before hand and to start from there on the interaction journey and to take and use notes during the whole process. We observed that both users and particularly intermediaries very often took notes during the interaction. They also used notes and graphical depiction in explanations to users. The nature of written notes during the searching process is an interesting research question, currently being investigated by Spink and Goodrum (1996). Furthermore, IR interfaces that accept a question in natural language and have the capability of accepting notes from a whiteboard may be more effective, particularly if coupled with components that suggest further or alternate terms as the interaction progresses.

Interestingly, as the number of Term Relevance Feedback terms in a question rose, the percentage of terms from User Interaction fell. It also rose with an increase in number of cycles, moves, and relevant terms. This shows some of the attributes of terms derived from relevance feedback: they tend to be used when users do not contribute many terms during interaction; by and large they have positive effectiveness and result in relevant terms, and they may be a prime illustration of interactivity (as reflected through cycles and moves). In other words, relevance feedback for the generation of search terms is a productive process. It should be encouraged in practice. And the capability should be made highly visible in IR interfaces.

These findings also have implications for the training of end-users and intermediary searchers. End-users should be encouraged to use terms from their domain knowledge and terms they identify in the retrieval output of the search for query formulation and to engage in various interactions. Talking to a librarian, search intermediary, or another person, before and during a database search, may stimulate the end-user to identify further search terms. They should be made aware of the practice and power of relevance feedback to find further terms.

Interaction Processes

We formulated a stratified model of IR interaction, which incorporates a surface, cognitive, situational, and affective level. We view the interaction as a set of episodes in which actions (‘‘things’’) happen on a given level, coupled with interplays between levels. We concentrated on manifestations and the behavior of search term selection as a process on the surface level. However, we also involved relevance judgments to assess the effectiveness of classes of terms, meaning that we also involved the cognitive and affective levels and possibly the situational level as well.

The distinction as to the levels served us well, because we were able to distinguish clearly between the manifestations where a relevance judgment was and was not incorporated. Thus, we present overall distributions of search terms (surface level), contrasted with distributions where effectiveness (derived from a cognitive judgment of relevance, thus surface plus cognitive level) is applied. We also included correlations with user satisfaction, which is on the affective level. An IR system deals with the surface level only, trying in a variety of ways to guess and simulate what may be effective on the other levels. But in interaction, it remains on the surface level. On the other hand, intermediaries deal on both surface and cognitive levels, and in their interaction with users often play a role where they affect the cognitive state of users above and beyond the IR system output. As yet, interfaces have not reached a high degree of capability to simulate the intermediaries in their interaction roles in general nor to play a significant role in selection of search terms in particular. Achieving such capabilities, even to a degree, should be the major goal of IR interface design.

As can be seen from basic data and from two instances where we presented data there were very large individual differences among questions. While we have not investigated the possible differences in interaction among questions as related to the selection of search terms, from casual perusal of the available data, we can see that indeed significant differences in interaction ways and means exist. This brings up a whole set of issues not as yet investigated, issues related to the nature, manifestations, and effects of different types, styles, and strategies in IR interactions. We have started such an investigation, as can be seen from the cited articles above, but there is a long way to go. Hopefully, the new and evolving models of IR interactions (including our own), and the ''intelligent'' interfaces on the horizon, will be able to accommodate, rather then sweep under the rug, such individual differences.

The selection of search terms for a question and the construction of queries is a highly interactive process. It is reiterative and not linear. While what goes on the surface level is highly important because the actions can affect the outcomes and effectiveness in many ways, the cognitive and situational dimensions are predominant for acceptance, use, and evaluation. The interaction with IR systems is still largely a human art. Mastery of such an art is teachable and can be improved by experience and practice. Enhancing features that will help master the art can be incorporated in teaching, using, and designing IR systems and processes. What is involved in such art is researchable and can be specified to some degree, as we and others have tried. It is not entirely a mystery. To design more effective, usable and acceptable interfaces, we have to learn from such art.

Acknowledgment

This work resulted from a grant (ref. no. R039A80026) from the Library Research and Demonstration Program, U.S. Department of Education, and a grant from DIALOG Information Services. The research would not be possible without the users and intermediaries. Their contribution is gratefully acknowledged, however anonymous. A number of students and faculty at the School of Communication, Information, and Library Studies, Rutgers University participated in design, conduct, critique, and analysis of the study from which we derived the data. Louise Su was the project director in charge of managing the collection of data. The contribution of all these colleagues is highly valued.

References

Bates, M. J. (1988). How to use controlled vocabularies more effectively in online searching. *Online, 11,* 45–56.

Bates, M. J., Wilde, D. N., & Siegfried, S. (1993). An analysis of search terminology used by humanities scholars: The Getty Online Searching Project Report Number 1. *Library Quarterly, 63,* 1–39.

Belkin, N. J. (1984). Cognitive models and information transfer. *Social Science Information Studies, 4,* 111–129.

Belkin, N. J. (1993). Interaction with texts: Information retrieval as information-seeking behavior. In *Information retrieval. 10. Von der Modelierung zur Anwerdung.* Konstanz, Germany: Universitaetsverlag. 55–66.

Belkin, N. J., Cool, C., Stein, A., & Thiel, U. (1995). Cases, scripts, and information seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications, 9*(3), 379–395.

Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM, 35*(12), 29–38.

Belkin, N. J., & Vickery, A. (1985). *Interaction in information systems.* London: The British Library.

Blair, D. C. (1990). *Language and representation in information retrieval.* New York: Elsevier.

Blair, D. C., & Maron, (1985). An evaluation of retrieval effectiveness for a full-text document-retrieving system. *Communications of the ACM, 28,* 289–299.

Cleverdon, C. (1967). *Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems.* Cranfield, England: College of Aeronautics, Aslib Cranfield Research Project.

Corbin, J., & Strauss, A. (1990). *Basics of qualitative research: Grounded theory procedures and techniques.* Newbury Park, CA: Sage Publications.

Croft, W. B. (1987). Approaches to intelligent information retrieval. *Information Processing and Management, 23,* 249–254.

Dubois, C. P. R. (1987). Free text vs. controlled vocabulary: A reassessment. *Online Review, 11,* 243-253.

Efthimiadis, E. (1993). A user-centered evaluation of ranking algorithms for interactive query expansion. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 16,* 146–159.

Fidel, R. (1986). Toward expert systems for the selection of search keys. *Journal of the American Society for Information Science, 37*(1), 37–44.

Fidel, R. (1991a). Searchers' selection of search keys. I. The selection routine. *Journal of the American Society for Information Science, 42*(7), 490–500.

Fidel, R. (1991b). Searchers' selection of search keys. II. Controlled vocabulary or free-text searching. *Journal of the American Society for Information Science, 42*(7), 501–514.

Fidel, R. (1991c). Searchers' selection of search keys. III. Searching styles. *Journal of the American Society for Information Science, 42*(7), 515–527.

Harman, D. (Ed.) (1995). The second Text Retrieval Conference—TREC 2. *Information Processing and Management, 31(3),* Special Issue, 244–448.

Harter, S. P. (1986). *Online information retrieval: Concepts, principles and techniques.* New York: Academic Press.

Hsieh-Yee, J. (1993). Effect of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science, 44(3),* 161–174.

Ingwersen, P. (1992). *Information retrieval interaction.* London: Taylor Graham.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation, 52(1),* 3–50.

Keen, E. M. (1973). The Aberystwyth Index Language Test. *Journal of Documentation, 29,* 1–35.

Knoke, D., & Burke, P. J. (1980). *Log-linear models.* Beverly Hills, CA: Sage Publications.

Lancaster, F. W. (1980). Trends in subject indexing from 1957 to 2000.

In P. J. Taylor (Ed.), *New trends in documentation and information.* London: Aslib.

Markey, K., et al. (1982). An analysis of controlled vocabulary and free text search statements in online searches. *Online Review, 4,* 225–235.

Mokros, H. (1984). *Patterns of persistence and change in the sequence of nonverbal actions.* Unpublished Ph.D. dissertation. University of Chicago.

Mokros, H., Mullins, L., & Saracevic, T. (1995). Practice and personhood in professional interaction: Social identity and information needs. *Library and Information Science Research, 17,* 237–257.

Parker, J. E. (1971). Preliminary assessment of the comparative efficiencies of an SDI system using controlled vocabulary or natural language for retrieval. *Program, 5,* 26–34.

Robertson, S. J., Walker, S., & Hancock-Beaulieu, M. M. (1995). Large text collection experiments on an operational interactive system: Okapi at TREC. *Information Processing and Management, 31(3),* 345–361.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART retrieval system: Experiments in automatic document processing.* Englewood Cliffs, NJ: Prentice-Hall, 313–323.

Rowley, J. (1994). The controlled versus natural indexing language debate revisited: A perspective on information retrieval practice and research. *Journal of Information Science, 20(2),* 108–119.

Salton, G., & Buckley, C. (1990). Improving information retrieval by relevance feedback. *Journal of the American Society for Information Science, 41(4),* 288–297.

Saracevic, T. (1975). Relevance: A review and a framework for the thinking on the notion of relevance. *Journal of the American Society for Information Science, 26(6),* 321–343.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development in Information Retrieval,* 138–146.

Saracevic, T. (1996). Interactive models in information retrieval (IR): A review and proposal. *Proceedings of the 59th Annual Meeting of the American Society of Information Science, 33,* 3–9.

Saracevic, T., Kantor. P., Chamis, A. Y., & Trivison, D. (1988). A study of information seeking and retrieving. I. Background and methodology. II. Users, questions and effectiveness. III. Searchers, searches and overlap. *Journal of the American Society for Information Science, 39(3),* 161–216.

Saracevic, T., Mokros, H., & Su, L. (1990). Nature of interaction between users and intermediaries in online searching: A qualitative analysis. *Proceedings of the 53rd Annual Meeting of the American Society for Information Science, 27,* 47–54.

Saracevic, T., Mokros, H., Su, L., & Spink, A. (1991). Interaction between users and intermediaries in online searching. *Proceedings of the 12th Annual National Online Meeting, 12,* 329–341.

Saracevic, T., & Su, L. (1989). Modeling and measuring the user-intermediary-computer interaction in online searching: Design of a study. *Proceedings of the 52nd Annual Meeting of the American Society for Information Science, 26,* 75–80.

Spink, A. (1993a). The effect of user characteristics on search outcome in mediated online searching. *Online & CD-ROM Review, 17(5),* 275–278.

Spink, A. (1993b). *Feedback in information retrieval.* Unpublished PhD dissertation. Rutgers University, School of Communication, Information and Library Studies.

Spink, A. (1993c). Interaction with information retrieval systems: Reflections on feedback. *Proceedings of the 56th Annual Meeting of the American Society for Information Science, 30,* 115–121.

Spink, A. (1994). Term relevance feedback and query expansion: Relation to design. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 17,* 81–90.

Spink, A. (1995). Term relevance feedback and mediated database searching: Implications for information retrieval practice and systems design. *Information Processing and Management, 31(2),* 162–171.

Spink, A. (in press). A feedback model for information retrieval. *Journal of the American Society for Information Science.*

Spink, A., & Goodrum, A. (1996). A study of search intermediary working notes. *Information Processing and Management, 32*(6), 681–696.

Spink, A., & Losee, R. M. (1996). Feedback in information retrieval. In Williams, M. (ed.) *Annual Review of Information Science and Technology, 31,* 33–78.

Spink, A., & Saracevic, T. (1992a). Sources and uses of search terminology in mediated online searching. *Proceedings of the 55th Annual Meeting of the American Society for Information Science, 29,* 249–255.

Spink, A., & Saracevic, T. (1992b). Where do the search terms come from? *Proceedings of the 13th Annual National Online Meeting, 13,* 363–373.

Spink, A., & Saracevic, T. (1993a). Dynamics of search term selection during mediated online searching. *Proceedings of the 56th Annual Meeting of the American Society for Information Science, 30,* 63–72.

Spink, A., & Saracevic, T. (1993b). Search term selection during mediated online searching. *Proceedings of the 14th National Online Meeting, 14,* 387–397.

Storrs, (1994). A conceptualization of multiparty interaction. *Interaction with Computers, 6(2),* 173–189.

Strzalkowski, T. (1995). Natural language information retrieval. *Information Processing and Management, 31(3),* 395–416.

Swanson, D.R. (1977). Information retrieval as a trial and error process. *Library Quarterly, 47*(2), 128–148.

Williams, F. (1992). *Reasoning with statistics: How to read quantitative research.* Fort Worth: Harcourt Brace Jovanovich.

Wu, M. M. (1992). *Information interaction dialogue: A study of patron elicitation in IR interaction.* Unpublished Ph.D. dissertation. Rutgers University. New Brunswick, NJ.