A Study of Information Seeking and Retrieving. III. Searchers, Searches, and Overlap*

Tefko Saracevic
School of Communication, Information and Library Studies, Rutgers, The State University of New Jersey, 4 Huntington St. New Brunswick, N. J. 08903

Paul Kantor
Tantalus Inc. and Department of Operations Research, Weatherhead School of Management, Case Western Reserve University, Cleveland, Ohio 44106

The objectives of the study were to conduct a series of observations and experiments under as real-life situation as possible related to: (1) user context of questions in information retrieval; (2) the structure and classification of questions; (3) cognitive traits and decision making of searchers; and (4) different searches of the same question. The study is presented in three parts: Part I presents the background of the study and describes the models, measures, methods, procedures and statistical analyses used. Part II is devoted to results related to users, questions and effectiveness measures, and Part III to results related to searchers, searches and overlap studies. A concluding summary of all results is presented in Part III.

Introduction

This is a third and concluding article on a study whose aim was to contribute to a formal, scientific characterization of the elements involved in information searching and retrieving, particularly in relation to the cognitive aspects and human decisions and interactions involved. The first part [1] presents the models, methods, measures, and procedures involved, together with a review of related works and all the background references. The second part [2] presents results on variables related to users, questions, and effectiveness measures. This third part is devoted to variables related to searchers and searching, and to overlap studies; it also contains conclusions for the study as a whole.

The second part contains as an introduction a summary of the objectives and procedures, so that a reader interested in results alone can proceed reading Parts II and III on their own. In addition, Part II includes an Appendix listing the questions used in the study together with data on items retrieved and evaluated. A Final Report [3] deposited with NTIS and ERIC describes all aspects of the study in great detail and includes a series of Appendices containing full question statements, "raw" data, forms used, and flowcharts of procedures.

Since Part II contains a summary of objectives and procedures we shall proceed directly with a presentation of the results. Part II includes Tables 1 to 18 and the table numbering continues here starting with Table 19.

Searchers

What Were the Results on Cognitive Tests?

The 39 searchers (36 outside and 3 project searchers) were tested on three cognitive tests (for references on these tests see Part I):

(1) **Remote Associates Test (RAT):** claims to be a test of semantic or word association. The scores can vary from 0 to 30.
(2) **Symbolic Reasoning Test (SRT):** claims to be a test of the ability to make deductive inference based on symbols. The scores can vary from 0 to 30.
(3) **Learning Style Inventory (LSI):** claims to determine an individual's preference for one of the postulated learning styles representing a characteristic method for acquiring and using information. The LSI yields six scores: one for each of the four learning modes (Concrete Experience (CE), Reflective Observation (RO), Abstract Conceptualization (AC) and Active Experimentation (AE)) and two for combination scores determining learning styles: AC − CE (the extent to which an individual emphasizes abstractness over concreteness in learning) and AE − RO (emphasizes action over reflection). The four learning mode scores

Work done under the NSF grant IST85-05411 and a DIALOG grant for search time.

Received July 9, 1987; accepted December 11, 1987.

TABLE 19. Searchers' scores for cognitive tests (*N* searchers = 39; for explanation of abbreviations see text).

Test	Mean	Standard Deviation	Min.	Max.
RAT	13.03	5.2	5	28
SRT	10.61	5.0	3	21
LSI:				
CE	24.70	7.3	13	42
RO	27.90	7.4	12	44
AC	33.55	9.4	16	48
AE	33.68	7.8	15	46
AC-CE	8.60	15.1	−20	35
AE-RO	5.73	13.3	−29	31

can vary from 12 to 48 and the two combination scores from−36 to +36.

In addition, searchers answered a question on their frequency of online searching on DIALOG (the service used in the study).

Table 19 provides the means, standard deviations, and ranges for the three cognitive tests taken by searchers. The distributions (not shown) were not normal: for RAT there was a peak at 10 and another at 18, for SRT a peak at 5 and another at 14; for LSI talking about peaks is not appropriate but by plotting the AC − CE scores against AE − RO scores in a graph the respondents are placed in categories as to learning styles: 16 (41%) of searchers were placed in the category called "converger," 2 (5%) in category "diverger," 10 (26%) in category "assimilator," 9 (23%) in category "accommodator," and 2 (5%) were "indeterminate."

As to the frequency of searching DIALOG, 12 (31%) of searchers reported searching DIALOG daily; 13 (33%) twice a week; 3 (8%) once a week; 2 (5%) twice a month; and 9 (23%) less than twice a month. Thus, 72% of searchers used DIALOG at least once a week.

What Was the Relationship between Searchers Characteristics and Relevance Odds?

The following question was asked: What were the odds that retrieved items be relevant or partially relevant (as opposed to not relevant) in searches done by searchers who scored high (above mean) on a given cognitive test as opposed to searchers who scored low (below mean)?

Table 20 provides the answers; it summarizes the relationship between relevance odds and a number of characteristics of searchers. The structure of this and all other tables on relevance odds was described at length with the presentation of Table 5 in Part II. Briefly, the columns starting from the left show the following: the cut point or mean that divides the high and low scores for the variable; the cross product ratio calculated from the 2 × 2 contingency table; the corresponding logarithm; standard error (+ or −); *t*-value used for test of significance; and the indication if the relation was significant at 95% (i.e., yes, when the magnitude of the *t*-value was above 2).

Scores from the 36 outside searchers only were used in this analysis, as well as the next one on precision and recall odds, because they did one search each, while the project searchers did more than one, thus with project searchers a learning factor may have been present.

The results indicate that items retrieved in searches done by searchers who:

- scored high (above mean) on a word association test (RAT) were 60% (or by a factor of 1.60) more likely to be relevant or partially relevant as opposed to not relevant;
- indicated that they preferred a Concrete Experience mode of learning were 29% (1–0.71) less likely to be relevant or partially relevant;
- indicated that they preferred an Abstract Conceptualization mode of learning style were 41% (or 1.41 times) more likely to be relevant or partially relevant;
- indicated that they emphasized abstractness over concreteness (AC-CE) as their learning style were 41% (or 1.41 times) more likely to be relevant or partially relevant.

In general, how a searcher scores on Remote Associates Test and what preference on learning he or she indicated affected relevance odds. Scores on Symbolic Reasoning Test and frequency of DIALOG searching did not have a significant effect.

The question of what these (and other) cognitive tests really represent, as well as the validity of their claims are

TABLE 20. Summary of the relation between searchers' characteristics and the odds that a retrieved item be relevant or partially relevant (Done for outside searchers only) (*N* outside searchers = 36; *N* outside searches = 200; *N* items retrieved = 4841; statistical significance at 95%; for abbreviations see text).

Searchers Characteristics	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/−	*t*-Value	Stat. Signif.
Frequency of searching	3.4	1.10	0.10	0.07	1.48	No
RAT (0 to 30)	13.03	1.60	0.47	0.06	7.84	Yes
SRT (0 to 30)	10.61	0.97	−0.03	0.06	−0.53	No
LSI:						
CE (12 to 48)	24.70	0.71	−0.34	0.06	−5.82	Yes
RO (12 to 48)	27.90	0.96	−0.04	0.06	−0.69	No
AC (12 to 48)	33.55	1.29	0.25	0.06	4.36	Yes
AE (12 to 48)	33.68	1.11	0.10	0.06	1.73	No
AC-CE (−36 to +36)	8.60	1.41	0.34	0.06	5.88	Yes
AE-RO (−36 to +36)	5.73	1.01	0.01	0.06	0.24	No

issues debated in psychology and other fields interested in human testing. We can express our doubts, but we cannot assert or question the validity of tests used here either way. Thus, taking the test claims strictly on their face value, it seems that searchers who show higher abilities or skills in language expressions, particularly word association, and/or searchers who lean toward abstractness in learning are more successful as to retrieval of relevant items. Surprisingly, searchers who score high on symbolic or logical reasoning had no significant effect on relevance odds; however, if we took all the searchers together (outside plus project searchers as we did in the analyses presented in the Final Report [3] and not only the outside searchers as we did here) then we observed a small negative correlation between high scores on Symbolic Reasoning Test and relevance odds.

While frequency of DIALOG searching had no significant effect on relevance odds, we have to underscore the fact that all searchers in the study were experienced and that no inexperienced searchers were included, thus this finding is not unexpected.

In practical terms, the findings on cognitive tests suggest that sharpening the semantic competencies in general and in a subject in particular is one of the more useful activities leading to possible improvement in searching, be that by human intermediaries or intelligent interfaces.

What Was the Relationship between Searchers Characteristics and Precision and Recall Odds?

The following question was asked: What were the odds that searches produced by searchers with above mean score on given characteristics had above mean precision or recall?

Table 21 provides the summary of the relationship between searchers' characteristics and precision and recall odds. (Abbreviations are explained at the beginning of this section.) The detailed explanation provided with Table 6 in Part II about the structure and contents of the tables on precision and recall odds are appropriate here as well. Briefly, column 1 provides the cut points or mean scores for given characteristic which divides the high and low values; column 2 gives the cross odds ratio representing the odds of moving to the high value of precision or recall due to a high value of a given characteristic; column 3 provides the associated log odds; column 4, the standard error; column 5, the *t*-value; and column 6, the indication of significance at 95%.

As can be seen, precision odds were significantly affected by only one characteristic (LSI-CE) and recall by three — all associated with the Learning Style Inventory. The results indicate that searches produced by searchers who indicated Concrete Experience as preferred learning mode were 50% (1–0.50) less likely to have high precision and 49% less likely to have high recall, or in opposite terms, they were two times (1/0.5 = 2) more likely to have low (below mean) precision and also 2 times as likely to have low (below mean) recall. Searchers who indicated Abstract Conceptualization as preferred mode of learning were 3.27 times more likely to have high (above mean) recall. No other characteristic, including frequency of searching had a significant effect on precision and recall.

In general, searchers indicating concrete experience as their preferred learning mode had significantly lower odds for both, precision and recall. Searchers preferring abstractness over concreteness in learning had improved recall odds.

This suggests that preference toward concreteness in learning diminishes both, precision and recall, while preference toward abstractness in learning enhances recall. Searchers who prefer abstract learning may have a better chance at higher precision and recall. These conclusions emphasize and amplify those made on the relation between searchers characteristics and relevance odds.

Searches

Results of searches are presented in this and the following two sections. In this section we treat the measures that reflect tactics and efficiency of searches. In the next section we concentrate on project searches and in the third section involving searches we deal with overlap among items retrieved by different searches for the same question, together with overlap in search terms. Although, search variables were involved in all three sections, we divided them to underscore the point that quite different research questions have been asked in each.

What Were the Figures for Tactics and Efficiency Measures?

Table 22 presents the measures related to tactics and efficiency for all searches (comparison between outside and project searches is given in Table 25). All measures are self explanatory with possible exception of command cycles: a cycle is a sequence of commands between beginning of a search and viewing of retrieved results (e.g. typing, printing) or between two viewings. Use of more than one cycle may indicate testing of output and possible adjustment of search strategy. The first three measures (number of commands, cycles, and search terms) relate to search tactics, while the last three (preparation, online, and total time) relate to efficiency or costs.

As can be seen, a typical search (if there is such a thing) had about 15 commands, 3 cycles and 10 search terms, took about 13 minutes of preparation time, 14 minutes of online time, for a total of 27 minutes from start to end. However, these mean figures, as all others, have to be interpreted with caution, because the ranges were wide and the distributions were not normal, they were all skewed toward the low end of the scales.

What Was the Relationship between Tactics and Efficiency Measures and Relevance Odds?

The following question was asked: What were the odds that searches with high (above mean) scores on tactics and efficiency measures retrieved items that are relevant or partially relevant as opposed to not relevant? The answers can

TABLE 21. Summary of the relation between searchers' characteristics and the odds that precision and recall be above average (Done for outside searchers only) (*N* outside searchers = 36; *N* searches = 200; statistical significance at 95%; mean precision for outside searches = 0.54; mean recall for outside searches = 0.20; for abbreviations see text).

Searchers Characteristic	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/−	*t*-Value	Stat. Signif.
Precision						
Frequency of searching	2.93	1.54	0.43	0.33	1.30	No
RAT (0 to 30)	13.03	1.44	0.36	0.28	1.27	No
SRT (0 to 30)	10.60	1.39	0.33	0.28	1.14	No
LSI:						
CE (12 to 48)	24.70	0.50	−0.69	0.29	−2.39	Yes
RO (12 to 48)	27.90	1.13	0.13	0.29	0.43	No
AC (12 to 48)	33.55	1.60	0.47	0.29	1.59	No
AE (12 to 48)	33.68	1.09	0.08	0.29	0.29	No
AC-CE (−36 to +36)	8.60	1.95	0.66	0.29	2.29	Yes
AE-RO (−36 to +36)	5.73	0.89	−0.12	0.28	−0.42	No
Recall						
Frequency of searching	2.93	0.70	−0.36	0.33	−1.09	No
RAT	13.03	1.10	0.10	0.29	0.33	No
SRT	10.60	1.09	0.09	0.29	0.29	No
LSI:						
CE	24.70	0.51	−0.68	0.29	−2.29	Yes
RO	27.90	1.39	0.33	0.30	1.08	No
AC	33.55	3.27	1.19	0.30	3.87	Yes
AE	33.68	0.62	−0.48	0.30	−1.62	No
AC-CE	8.60	2.47	0.90	0.30	3.01	Yes
AE-RO	5.73	0.55	−0.59	0.29	−2.00	No

TABLE 22. Tactics and efficiency measures for all searches (*N* searches = 360).

For a Search:	Mean	Standard Deviation	Min.	Max.
No. of Commands	14.5	7.7	2	50
No. of Command Cycles	3.4	2.0	1	14
No. of Search Terms	10.3	7.0	1	61
Preparation Time (hrs.)	0.22	0.14	0.02	0.83
Online Connect Time	0.24	0.16	0.01	1.24
Total Time Used	0.46	0.25	0.10	1.96

be found in Table 23; this is a summary of relevance odds as related to the three tactics and three efficiency measures.

As can be seen, the number of commands in a search had no significant effect, but a high (above mean) number of command cycles increased relevance odds by a factor of 1.18. A high (above mean) number of search terms decreased relevance odds by 22% (1–0.78); or in opposite terms: items retrieved in searches with a low number of search terms were 28% (1/0.78) more likely to be relevant. High preparation time and high total time used for a search decreased relevance odds by 13% and 19% respectively; or in opposite terms: items retrieved in searches with low preparation time and total time were 15% (1/0.87) and 23% (1/0.81) respectively more likely to be relevant.

It may be of interest to comment on two seemingly contradictory findings on relevance odds. We report here that a high number of search terms in a search decreased relevance odds, while earlier (Part II, Table 10) we reported that a high degree of complexity in a question (high number of search concepts) increased relevance odds. The two

TABLE 23. Summary of the relation between tactics and efficiency measures for searches and odds that a retrieved item be relevant or partially relevant (*N* all searches = 360; *N* items retrieved = 8956; statistical significance at 95%).

Tactics or Efficiency Measure	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/−	*t*-Value	Stat. Signif.
No. of Commands	14.51	0.94	−0.07	0.04	−1.51	No
No. of Command cycles	3.40	1.18	0.17	0.04	3.84	Yes
No. of Search Terms	10.33	0.78	−0.25	0.05	−5.42	Yes
Preparation Time (hrs)	0.22	0.87	−0.14	0.04	−3.31	Yes
Online Connect Time	0.24	1.08	0.08	0.04	1.71	No
Total Time Used	0.46	0.81	−0.21	0.04	−4.73	Yes

may not be the same. Questions of high complexity may have searches with a low number of search terms and vice versa. However, the issue is an interesting one for further research.

In general, none of the effects of tactics and efficiency measures on relevance odds were large. The only positive effect was number of command cycles, while number of commands had no significant effect. In contrast, the number of search terms, preparation time, and total time used in a search had a negative effect. Online connect time had no significant effect.

This suggests that the search tactics (as measured in this and similar studies reviewed in Part I) and search time have some, but not a large effect on relevance odds, particularly in comparison to many other variables tested. How a search is done seems to affect the results less than many other factors associated with users, questions, searchers, and search terms selections. More cycles, which allow for feedback, seem to produce better searches as to relevance odds, and more search terms, a lot of preparation time and consequently a lot of total time seems to result in worse searches. The finding on time is somewhat surprising. It seems to suggest that when searchers tend to use a lot of time they may not know what they are after in the first place. However, as for other speculations, this needs further research.

What Was the Relationship between Tactics and Efficiency Measures and Precision and Recall Odds?

The following question was asked: What were the odds that searches with above mean values on tactics and efficiency measures had above mean precision and recall? To answer the question directly: None of the measures had a significant effect on either precision or recall.

Table 24 provides the summary of the relation between the six measures used to indicate tactics and efficiency of searching and precision and recall odds. The table is arranged the same way as in the previous tables on precision and recall odds.

All of the measures turned to have no significant impact at 95%. The finding of no significant effect of tactics and efficiency measures on the macro or search-wise level of analysis is not surprising, given the fact that effects on the micro or item-wise level of analyses of the same variables were rather small. This underscores the suggestion that search tactics and search time have a rather small effect on search performance. How a search is executed, by itself, seems to be a small contributing factor to the success of the outcome. As some other findings in this study, this finding is a challenge to many accepted (but untested) models of searching (see Part I for review and references). Again, this assertion, as others, requires further research. In particular, the effects of a variety of searching rules for human searchers suggested in the literature should be investigated (something that was not done in this project). Which ones work and which ones do not? Which ones could be automated? How do they compare to existing machine searching rules? A whole research agenda could follow.

Project Searches

In relation to project searches the objective was to investigate the effect of searches based on sources in addition to the written question. To construct a search, the outside searchers were given the written question exactly as presented by the user. They were also given the thesauri and other tools appropriate for the files to use as they found necessary. The project searchers were asked to construct four types of searches based on:

Type 1. The taped statement of users describing the problem at hand and the intent in use of information.

TABLE 24. Summary of the relation between tactics and efficiency measures for searches and the odds that precision and recall be above average (*N* searches = 360; statistical significance at 95%; mean precision for all searches = 0.57; mean recall for all searches = 0.22).

Variable	Cut Point (Mean)	Odds Ratio	Log Odds	Stand. Error +/−	*t*-Value	Stat. Signif.
Precision						
No. of commands	14.51	1.09	0.09	0.22	0.40	No
No. of command cycles	3.40	1.39	0.33	0.22	1.51	No
No. of search terms	10.33	0.85	−0.17	0.22	−0.76	No
Preparation time (hr)	0.22	0.75	−0.29	0.21	−1.36	No
Online connect time	0.24	1.07	0.07	0.22	0.30	No
Total time used	0.46	0.96	−0.04	0.22	−0.17	No
Recall						
No. of commands	14.51	0.95	−0.05	0.22	−0.23	No
No. of command cycles	3.40	1.04	0.04	0.22	0.17	No
No. of search terms	10.33	0.89	−0.11	0.22	−0.50	No
Preparation time (hr)	0.22	0.91	−0.09	0.22	−0.42	No
Online connect time	0.24	1.07	0.07	0.22	0.31	No
Total time used	0.46	0.91	−0.10	0.22	−0.44	No

Type 2. The taped problem statement *plus* the written question.
Type 3. The terms extracted from the written question *without* any elaboration (as if performed by automatic extraction of keywords).
Type 4. The written question *plus* elaboration from a thesaurus.

As mentioned, there were 5 outside searches per question and 4 project searches, for a total of 9 per question, or for the 40 questions there were 200 outside and 160 project searches, for a total of 360 searches.

What Does a Comparison between Outside and Project Searches Show?

Table 25 presents a comparison of means and standard deviations between outside and project searches on (i) retrieved items, (ii) precision and recall, and (iii) tactics and efficiency measures.

There are some differences between outside and project searches but they were relatively small. On the average, per search, the project searches tended to:

- produce about 2 more relevant and 2 more partially relevant items, and 1 less not relevant item;
- have a precision that is higher by 7 percentage points and recall that is higher by 5 points;
- use about 3 fewer commands, 1 less cycle, and the same number of search terms;
- take about 4 minutes less to prepare, 4 minutes less of online time, and 8 minutes less from start to end.

On the average, the project searchers have higher performance figures than the outside searchers. This is not surprising since the project searchers used additional bases for search construction, they did four searches for the same question and there may have been a learning factor involved, while the outside searchers did only one search per question.

In the analysis of other variables we have used as appropriate for the particular research question either the 200 outside searches or the 360 outside plus project searches. This was done in order to eliminate (where necessary) the learning factor. If we use either all searches (outside + project) or only outside searches in all analyses, the results may change a few percentage points, but our conclusions would remain the same. In other words, no conclusion as to direction of relevance, precision and recall odds would change by using either outside or all searches and even the magnitudes would be very close.

What Does a Comparison Between Four Types of Project Searches Show?

The possibility of a distinction among the four types of project searches was studied by analysis of variance applied to precision and recall. The cross product ratio method was not applied, because there were four different classes of valuables, while the method is limited to two binary variables.

Table 26 provided means and standard deviations for precision and recall respectively for the four types of searches. We see that:

- searches of *type 1* (problem statement) have the highest mean precision (about 64%) and also the highest recall (32%);
- searches of *type 4* (question plus thesaurus) were third best in mean precision (61%) and second best in mean recall (25%);
- searches of *type 2* (problem statement plus question) were second best in mean precision (63%), but third best in mean recall (23%);
- searches of *type 3* (written question only) had the lowest mean precision (57%) together with the lowest mean recall (18%).

The analysis of variance comparing recall for the four types of project searches reveals a significant difference. A corresponding analysis for precision reveals no significant difference. In other words, different types of searches had

TABLE 25. Comparison of statistics on outside vs. project searches (*N* outside searches = 200; *N* project searches = 160).

	Outside searches		Project searches	
Variable	Mean	Stand. Dev.	Mean	Stand. Dev.
Relevant items	6.89	9.71	8.57	11.89
Partially relevant	6.63	9.79	7.58	12.19
Not relevant	10.68	13.76	9.57	14.06
Total evaluated	24.21	23.90	25.72	25.36
Retrieved but not evaluated	25.03	62.66	23.40	42.31
Total retrieved	49.24	77.35	49.12	60.55
Search precision	0.54	0.35	0.61	0.32
Search recall	0.20	0.20	0.25	0.23
Commands used	15.73	8.13	12.99	6.94
Command cycles	3.73	2.25	2.98	1.50
Search terms	10.22	7.09	10.47	6.97
Preparation time (hrs.)	0.25	0.16	0.18	0.11
Online time	0.26	0.18	0.20	0.12
Total time	0.51	0.29	0.38	0.17

TABLE 26. Precision and recall for four types of project searches (*N* project searches of each type = 40). *Note:* There is no significant difference for precision at 95% (or even at 90%) significance, however, there is a significant difference between search types on recall at 95% significance. Thus, the table is arranged from highest to lowest recall.

Search Type and Description of Source for Search Terms	Precision		Recall	
	Mean	Stand. Dev.	Mean	Stand. Dev.
TYPE 1 Taped problem statement only	0.64	0.31	0.32	0.27
TYPE 4 Written question plus thesaurus elaboration	0.63	0.32	0.25	0.25
TYPE 2 Taped problem statement plus written question	0.57	0.32	0.23	0.19
TYPE 3 Restricted as found in written question only	0.61	0.34	0.18	0.16
ALL TYPES combined	0.61	0.32	0.25	0.23

significantly different recall, but not significantly different precision.

We did an analysis of variance on every variable where appropriate in the study, but this was the only case in this study where significant effects in analysis of variance were found in relation to precision and recall. All other variables showed no effect at this level of analysis, further underscoring the points made in conclusion of section on "Precision and Recall" in Part II.

In general, searches based on problem and intent statements by users out-performed on the average all other types of searches, including outside searches based on written questions. Searches done on the basis of terms from questions only (without elaboration) performed the poorest. Interestingly, when searches were based on problem statement plus a written question they did somewhat more poorly than searches based on problem statement alone. This suggests that the users' context, (the problem at hand and the intent) is a most powerful element in the potential effect on retrieval effectiveness, and that exploring the context has a large potential payoff, while doing the search on the basis of question terms only (without elaboration) is the poorest way to go about it. Automating the search process by only taking keywords from a question as search term may be a poor way of searching. The implications for research, systems design, and practice are obvious. One should be careful in searching not to rely on exclusively on words of a written question.

Overlap Studies

Two overlap aspects were studied for searches of the same written question: the degree of agreement in (i) selection of search terms and (ii) items retrieved. Overlap was computed for each pair of searchers and their search. The overlap measure between search 1 (S_1) and search 2 (S_2) is asymmetric:

$$S_{1,2} = \frac{|S_1 \cap S_2|}{|S_1|}; \qquad S_{2,1} = \frac{|S_1 \cap S_2|}{|S_2|}$$

In other words, the overlap between search 1 and 2, taking results of search 1 as the base, is equal to the number of search terms (or items retrieved) in common divided by the number of search terms (or items retrieved) by search 1. The overlap between search 2 and 1, taking search 2 as the base, is divided by the number of search terms (or items retrieved) by search 2.

As mentioned, each of the 40 questions was searched by 5 outside and 4 project searches. The overlap for a question is calculated only for the outside searches because they were based on the same written question by users. The project searches were based on additional sources and thus not comparable for our objective. There were 5 outside searches per question and for each search there were 4 comparisons (it was not compared with itself); thus there were 20 pairs of comparisons for each question. For 40 questions there were altogether 800 pairs of comparisons (5 × 4 × 40). The results are based on 800 overlap measurements.

What Was the Overlap in Search Terms?

Table 27 provides a distribution of the values of search terms overlap for the 800 pairs of comparisons. The table is read as follows: in 89 or 11.1% of cases the overlap was between 0.0 and 0.5 or 0% and 5%. In 70 or 8.8% of cases

TABLE 27. Agreement between outside searchers in selection of search terms (*N* outside searches per question = 5; *N* questions = 40; *N* pairs of comparisons = 800 = 5 searches per question *X* 4 comparisons for each *X* 40 questions.) Mean degree of agreement = 0.27; standard deviation = 0.20. Each row indicates a 5% range of agreement.

Degree of Agreement: at Least but Not Over	Frequency in Each Range	% of Total	Cumulative Percentage
0.0 to 0.05	89	11.1%	11.1%
0.05 to 0.10	70	8.8	19.9
0.15	82	10.3	30.2
0.20	113	14.1	44.3
0.25	97	12.1	56.4
0.30	52	6.5	62.9
0.35	72	9.0	71.9
0.40	33	4.1	76.0
0.45	47	5.9	81.9
0.50	6	0.8	82.7
0.55	69	8.6	91.3
0.60	24	3.0	94.3
0.65	8	1.0	95.3
0.70	10	1.2	96.5
0.75	13	1.6	98.1
0.80	3	0.4	98.5
0.85	0	0.0	98.5
0.90	0	0.0	98.5
0.95	0	0.0	98.5
1.00	12	1.5	100.0
TOTALS	800	100%	

the overlap was between 5% and 10%, and in 82 or 10.3% of cases it was between 10% and 15% (after the second row the first number in the range is not repeated, but it is still there). When the first five figures are cumulated, we can see that in 56.4% (or 451) of cases the agreement was less than 25% (i.e. between 0% and 25%). The mean overlap was .27 with a standard deviation of .20.

In general, the overlap in selection of search terms by different searchers searching the same question is relatively low. Given the same question, different searchers tend to select a few terms that are the same, and a considerably larger number that are different.

What Was the Overlap in Items Retrieved?

Table 28 provides a distribution of values for overlap of all retrieved items (regardless if judged relevant, partially relevant or not relevant) and for overlap of items judged relevant or partially relevant only for the 800 pairs of comparison. The table is read in the same way as the previous one. We can see that in 469 cases (58.6%) of all items retrieved and in 471 cases (58.9%) of relevant or partially relevant items retrieved the degree of overlap was between 0.0 and 0.05 (or 0% and 5%). When the first four figures are cumulated, we can see that in 75.3% (or 602) of cases involving retrieval of all items the overlap was less than 0.20 (or between 0.0 and 0.20) and it was in the same range of 72.8% (or 583) of cases involving relevant or partially relevant items retrieved. The mean overlap for all items and relevant or partially relevant items retrieved was 0.17 and 0.18 respectively (or 17% and 18%) with standard deviation of 0.28 and 0.30.

In general, the overlap in retrieved items (be they all items or relevant or partially relevant items only) by different searchers searching the same question is also relatively low, in fact it is significantly lower than the overlap in search terms by the same searchers. It seems that different searchers for the same question more or less look for and retrieve a different portion of the file. They seem to see different things in a question and/or interpret them in a different way and as a result retrieve different items.

Does the Search Term Overlap Explain the Retrieved Items Overlap?

The short and surprising answer: it does not.

A search for a regression relation between the two variables was not successful. A regression analysis shows that only 2.5% of the variation in overlap of retrieved items can be attributed to the overlap in search terms. (The scatter plot of the relations on the basis of which the regression analysis was performed is reproduced in the Final Report [3].)

In general, in searches for the same question by different searchers, the overlap in search terms and the overlap in items retrieved are not closely related. This further underscores the conclusion that different searchers for the same question see and interpret different things in a question, represent them by different linguistic and/or logical constructs, and retrieve different things from a file.

TABLE 28. Agreement between outside searches on retrieval of all items (i.e., $R + pR + N$) and of relevant or partially relevant items (i.e., $R + pR$ only). (N outside searches per question = 5; N questions = 40; N pairs of comparisons = 800.) Mean degree of agreement for all items retrieved = 0.17; standard deviation = 0.28. Mean degree of agreement for relevant or partially relevant items only = 0.18; standard deviation = 0.30. Each row indicates a 5% range of agreement.

Degree of Agreement: At least but not over	All Items Retrieved ($R + pR + N$)			Relevant or Partially Relevant ($R + pR$ only)		
	Frequency in Each Range	% of Total	Cumulative Percentage	Frequency in Each Range	% of Total	Cumulative Percentage
0.0 to 0.05	469	58.6%	58.6%	471	58.9%	58.9%
0.05 to 0.10	62	7.8	66.4	40	5.0	63.9
0.15	35	4.4	70.8	49	6.1	70.0
0.20	36	4.5	75.3	23	2.8	72.8
0.25	27	3.4	78.7	21	2.6	75.4
0.30	10	1.2	79.9	13	1.6	77.0
0.35	19	2.4	82.3	25	3.1	80.1
0.40	11	1.4	83.7	12	1.5	81.6
0.45	9	1.1	84.8	14	1.8	83.4
0.50	7	0.9	85.7	3	0.4	83.8
0.55	19	2.4	88.1	27	3.4	87.2
0.60	11	1.4	89.5	9	1.1	88.3
0.65	6	0.7	90.2	4	0.5	88.8
0.70	15	1.9	92.1	13	1.6	90.4
0.75	4	0.5	92.6	5	0.6	91.0
0.80	5	0.6	93.2	5	0.6	91.6
0.85	8	1.0	94.2	8	1.0	92.6
0.90	3	0.4	94.6	6	0.8	93.4
0.95	3	0.4	95.0	2	0.3	93.7
1.00	41	5.0	100.0	50	6.3	100.0
TOTALS	800	100%		800	100%	

What Was the Relationship between Multiple Retrievals and Relevance Odds?

As mentioned, there were 9 searches per question (5 outside and 4 project searches). From Table 1 in Part II we can see that the total number of all items retrieved by all searches was 8,956 of which 5,411 were unique; for outside searches the total number of retrieved items was 4,841 of which 3,691 were unique; for project searches the total number was 4,115 of which 2,920 were unique. (Reminder: the total number of all items, *including* duplicates, is a sum of the total number of items retrieved by outside and project searches; however, the total number of unique items, *excluding* duplicates, is a union of the two.) The difference between the total (including duplicates) and unique (excluding duplicates) numbers of retrieved items represents multiple retrievals, i.e., items retrieved more than once. In items retrieved from all (9) searches per question there were 3,545 (8,956–5,411) multiple retrievals; outside searches had 1,150 (4,841–3,691) multiple retrievals and project searches had 1,195 (4,115–2,920) multiple retrievals.

The following question was asked: What are the odds that an item retrieved once, twice . . . n times for the same question by different searches be relevant? In other words, we are considering here the relevance odds of single and multiple retrievals.

We calculated the relevance odds separately for retrievals by all, outside and project searches, as presented in Tables 29, 30, and 31. Since there were 9 total searches, 5 outside and 4 project searches, the frequency of retrieval for the first (Table 29) varies from 1 to 9, for the second (Table 30) from 1 to 5, and for the third (Table 31) from 1 to 4. The tables provide two things: (i) distribution of items retrieved from 1 to n = 9, 5, 4, and (ii) corresponding relevance odds. However, this time in addition to calculating the odds as elsewhere, i.e., by comparing relevant or partially relevant ($R + pR$) items with those judged not relevant (N), we also calculated odds for relevant alone vs. partially relevant or not relevant ($pR + N$), and relevant items alone vs. not relevant alone, disregarding the partially relevant. The three corresponding odds are: (i) $(R + pR)/N$ (called “normal relevance”, (ii) $R/(pR + N)$ (“weak relevance”), and (iii) R/N (“strong relevance”). This was done to sharpen the insight into behavior of relevance. Of particular interest are the last odds (R/N, “strong relevance”) because they zero in on the subset most preferred by the users. (Before discussing the results note that the sum of frequency times unique retrievals equals the number of all retrievals, e.g., from Table 30: (1 × 2915) + (2 × 488) + (3 × 209) + (4 × 72) + (5 × 7) = 4,841. This is also valid for items judged R, pR, or N.)

TABLE 29. Multiple retrieval of items by all (outside + project) searches for the same question (N all searches per question = 9; N questions = 40; N all searches = 360; N all items retrieved = 8956; N unique items retrieved = 5411).

No. of times retrieved	Number of *R*	*pR*	*R* + *pR*	*N*	Tot.	Perc.	Odds *R* + *pR* vs. *N*	*R* vs. *pR* + *N*	*R* vs. *N*
1	770	945	1715	2020	3735	69%	0.85	0.26	0.38
2	245	270	515	355	870	16.1	1.45	0.39	0.69
3	103	70	173	128	301	5.6	1.35	0.52	0.80
4	67	61	128	68	196	3.6	1.88	0.52	0.98
5	82	54	136	24	160	3.0	5.67	1.05	3.41
6	42	19	61	14	75	1.4	4.36	1.27	3.00
7	24	20	44	9	53	1.0	5.72	0.85	3.09
8	8	6	14	2	16	0.3			
9	2	3	5	0	5	0.1			
Any (1 to 9)	1343	1448	2791	2620	5441	100%	1.06	0.33	0.51

TABLE 30. Multiple retrieval of items by outside searches for the same question (N outside searches per question = 5; N questions = 40; N total outside searches = 200; N total items retrieved for outside searches = 4,841; N unique items retrieved = 3,691).

No. of times retrieved	Number of *R*	*pR*	*R* + *pR*	*N*	Tot.	Perc.	Odds *R* + *pR* vs. *N*	*R* vs. *pR* + *N*	*R* vs. *N*
1	620	750	1370	1545	2915	79%	0.89	0.27	0.40
2	185	129	314	174	488	13.2	1.80	0.61	1.06
3	92	61	153	56	209	5.7	2.73	0.79	1.64
4	23	30	53	19	72	2.0	0.16	0.52	1.42
5	4	3	7	0	7	0.2			
Any (1 to 5)	924	973	1897	1794	3691	100%	1.06	0.33	0.52

TABLE 31. Multiple retrieval of items by project searches for the same question (N project searches per question = 4; N questions = 40; N total project searches = 160; N total items retrieved for project searches = 4115; N unique items retrieved = 2920).

No. of times retrieved	Number of *R*	*pR*	*R* + *pR*	*N*	Tot.	Perc.	Odds *R* + *pR* vs. *N*	*R* vs. *pR* + *N*	*R* vs. *N*
1	503	626	1129	990	2119	72.6%	1.14	0.31	0.51
2	155	147	302	184	486	16.6	1.64	0.47	0.84
3	130	60	190	46	236	8.1	4.13	1.23	2.83
4	42	28	70	9	79	2.7	7.77	1.13	4.67
Any (1 to 4)	830	861	1691	1229	2920	100%	1.38	0.40	0.67

As can be seen, about 69% of all items (R + pR + N) in all searches (79% in outside and 73% in project searches) were retrieved only once; 16 in all searches (13% in outside and 17% in project) were retrieved twice; 6% (6%, 8%) were retrieved three times; and 9% (2%, 3%) were retrieved four or more times. Of relevant or partially relevant items in all searches about 61% (72% in outside, 67% in project) were retrieved only once, 18% (17%, 18%) twice, 6% (8%, 11%) three times and, 14% (3%, 4%) four or more times. The picture changes when we consider the not relevant items: of all not relevant items retrieved by all searches 77% (86%, 80%) were retrieved only once: 14% (10%, 15%) twice; 5% (3%, 4%) three times and 4% (1%, 0.7%) four or more times.

Let us now consider relevance odds. In "normal" relevance, $((R + pR)/N)$ the odds that an item retrieved any number of times in all searches (i.e., 1 to 9 times) be relevant or partially relevant as opposed to not relevant were about even or about 10 to 10 (2791/2620 = 1.06) (*Table 29*). For items retrieved only once, the odds were about 8 to 10 (0.85). For items retrieved twice, the odds increased by a factor of 1.45—they were about 15 to 10. For items retrieved three times, the odds were about 14 to 10 and four times 19 to 10. For items retrieved 5 times, the relevance odds jumped to a dramatic 57 to 10, for 6 times to 44 to 10, and for 7, 8 or 9 times, to 57 to 10. Relevance odds jumped to six fold between items retrieved only once and items retrieved 5 or more times. In other words, an item retrieved 5 or more times (out of 9 searches) was about over six times as likely to be relevant or partially relevant as an item retrieved once, and about five times as many as an item retrieved any number (1 to 9) times.

When considering "weak relevance" $(R/(pR + N))$, and "strong relevance" (R/N), the odds were, of course, smaller. However the order of magnitude in increase between items retrieved once, or any number of times, and items retrieved 5 or more times was not dissimilar to the magnitude observed above for "normal relevance." For "weak relevance," the odds for items retrieved 5 or more times were about 4 times higher than for items retrieved once, and 3 times higher than for items retrieved any number of times. For "strong relevance" the odds for items retrieved 5 or more times were about 8 times higher than for items retrieved once, and 6 times higher than for items retrieved any number of times. Thus, whether considering "normal," "weak" or "strong" relevance, we have obtained similar differences between relevance odds of multiple retrievals and single or any retrievals. No matter how we calculated relevance odds, the picture of differences in terms of order of magnitude remained similar. Consequently, to simplify matters in our conclusions from now on, we are discussing "normal relevance" only.

As expected, the results from outside searchers were not very different (Table 30). The odds of an item being relevant or partially relevant as against not relevant, $((R + pR)/N)$, for any number of retrievals (1 to 5) were about even (1896/1793 = 1.06). For items retrieved only once, they were about 9 to 10; for those retrieved twice, they were 18 to 10; for those retrieved 3 times, they improved to 27 to 10; and for those retrieved 4 or 5 times, the relevance odds jumped to a 33 to 10. For project searches (Table 31) the relevance odds for any number of retrievals (1 to 4) were about 14 to 10. For items retrieved only once they were 11 to 10; for those retrieved twice, 16 to 10; for those retrieved 3 times, they improved 40 to 10; and for those retrieved 4 times, they leaped to 78 to 10.

Considering any of the three cases (all searches, outside searches or project searches) for a question, the more often an item was retrieved the more the odds shifted in favor of relevance. We consider this as one of the most important findings of the study.

In general, the overlap in retrieved items for different searches for the same question done by different searchers is relatively low. However, the chances for relevance improved dramatically in items retrieved by more than one searcher. To underscore, the more often an item was retrieved by different searches for the same question the more likely it was to be relevant.

A Generalization on Overlap

A number of overlap or degree of agreement studies on a variety of processes associated with information representing, seeking and retrieving have been reported in the literature; a representative sample is enumerated below, others are reviewed by Bates [4]. The studies addressed overlap either in human decisions (e.g., consistency in indexing) or in systems performance (e.g., retrieval from different representations). These studies are, of course, related to a large family of investigations in psychology and cognitive science on human variability in forming associations and concepts, and in naming. At their basic level all such studies investigate differences in patterns formed by human minds. Since a majority of overlap studies on information representing, seeking, and retrieving have reported results in an equivalent range, a generalization may be in order.

Our results on the overlap in selection of search terms by searchers are in the same range, if not somewhat lower, as the results in selection of index terms by indexers, found in many inter-indexing consistency studies done during the 1960s and early 1970s, (e.g., see review and results by Zunde and Dexter [5]). Interestingly, they correspond also with overlap results on Dewey Decimal Classification assignments in online catalog retrieval reported by Markey and Demeyer [6], and in comparing overlap and searching behavior in a card catalog and an online public access catalog by Dalrymple [7].

The overlap in retrieval of items by different searches for the same question as found here is comparable to findings in various recent studies that observed overlap in retrieval between: different document representations, by Katzer *et al.* [8], keywords index terms, or descriptors and citations by Pao [9], and Pao and Fu [10], and Pao and Worthen [11], document sets clustered by co-citation and Mesh terms, by Rapp [12], and descriptors and citations, by McCain [13].

A general conclusion may be stated: the degree of agreement or overlap in human decisions related to representing, searching, and retrieving of information is relatively low—the agreement hardly reaches about one fourth or one third of the cases involved. However, the notion of "low" here may not be appropriate. We do not know what is "high" and the observed ranges may be all that is to be expected, i.e., they may be "normal." But whatever the findings on the range of human information behavior in these processes may be labeled, the results have a potential for a large impact on the direction and choices in research, design and practice, as suggested below.

Of course, all this calls for explanation and further research. A number of studies in cognitive science and artificial intelligence have been devoted to learning something about how human intelligence finds (or imposes) clusters and concepts on a set of patterns, (e.g. some ingenious ways of studying this are reported by Stepp and Michalski [14]). We may interpret our overlap results in terms of an opposite concept, namely in terms of "de-clustering". A structure (in our case a question) is decomposed or de-clustered by different searchers into a set of alternative forms (in our case searches). While the files which are searched are structured (indexed, organized) in a way that aims toward a clustering of answers, searchers seem to be working in a way that de-clusters them. This is an intriguing hypothesis and we are following it in our next phase of investigation.

As noted by Bates [4], the present design of online subject access, be it through library catalogs or online retrieval systems, does not accomodate human variability in searching. She calls (along with many other workers on information systems design [e.g., ref. 15]) for radically different design principles and implementations in order to accomodate the observed patterns, interactions and differences in human information behavior, of which the overlap findings are one of the important manifestations. A number of such desirable design features is enumerated in [4] and [15].

Let us end the discussion of overlap findings with some practical implications. As mentioned, although searchers disagree substantially in the items they retrieve in searching the same question, when they do agree they are likely to be producing relevant items. This suggests further that one possible super-strategy for the conduct of an online search is to have several searchers search a question independently and then to examine first the intersection (overlap) of their retrieved sets. The odds for finding relevant items in such an intersection are dramatically higher than in individual searches as a whole.

Conclusions

This study addresses cognitive decisions and human-systems interaction involved in information seeking and retrieving. The objectives were to conduct a series of observations or experiments under as real-life conditions as possible on five classes of variables: users, questions, searchers, searches, and items retrieved. Effects were explored on two levels: (i) micro or item-wise level concentrating on analysis of the relation between relevance of items retrieved and given variables, and (ii) macro or search-wise level concentrating on analysis of the relation between precision and recall of searches and given variables. In addition to standard statistical tools (distribution, analysis of variance, regression and correlations), the analysis of relations involved a powerful method called cross product ratio analysis, providing odds that relate a given variable with relevance of items retrieved, or with precision and recall of searches.

To our knowledge this is the largest study of its kind conducted to date. It involved a large amount of data and great many correlations of statistical significance. As a matter of fact, every meaningful combination of variables was explored for correlation on both levels, search-wise and item-wise. As expected, for many of these correlations we did not find statistically significant results. In the presentation of results we have highlighted the results that produced significant correlations while by and large ignoring those variables and combinations that did not.

There is, of course, a limit to our conclusions. As mentioned in Part 1, we cannot really claim generalizations beyond our sample any more than any other similar study has been able to claim. Still, we are offering these general conclusions to be taken with all due caution for discussion, confirmation, refutation, or (which is our best hope) as hypotheses for further study.

The conclusions that are boldfaced are specific to the variables and major findings in the study. Others are more general and they pertain to the models used, explanations, and suggestions for future studies.

Summary of Relevance, Precision and Recall Odds

Under the discussion of each class of variables (users, questions, searchers, searches, items retrieved), we presented separate tables on relevance odds and precision and recall odds for the given variable. We extracted from these tables the main conclusions to present together relevance, precision, and recall odds for all variables tested, as shown in Table 32. In effect, this table is a summary of all the findings on odds as related to the variables in this study. It is the highlight of the study. The stress is on description of the characteristic of the variables that *increased* odds that a retrieved item be relevant or partially relevant or that a search be of high (above mean) precision or recall.

In other words, we are restricting the conclusions in Table 32 to a narrative description of conditions that enhance (increase) the corresponding odds by a given amount, expressed as a factor of increase. We are *not* describing here the opposite conditions that decreased relevance, precision, and/or recall odds; however, these could be easily derived by using the explanation about odds interpretation presented throughout the article. Where the odds were not significant, we describe the variable in neutral terms, indicating that whether the condition be high or low the odds were not significantly effected.

The following questions were asked: What were the odds that items retrieved in association with a given characteristic of a variable be relevant or partially relevant as opposed to not relevant? What were the odds that searches associated with a given characteristic of a variable be of high (above mean) precision or recall, as opposed to low precision or recall? The interpretations from Table 32 are presented in boldfaced conclusions that follow.

As before, the reader is cautioned that interaction effects involving several variables together have *not* been ana-

TABLE 32. Summary of the impact of all variables in the study (for which such measurement could be made) on three measures of retrieval effectiveness: odds that retrieved item be relevant or partially relevant as opposed to not relevant; odds that a search had above average (high) precision and above average (high) recall (N.S. = not significant; note that recall odds are comparative only).

VARIABLE	Relevance	Precision	Recall
When the following	Odds	Odds	Odds
happened . . .	Increased by a factor of		
User Context			
Underlying PROBLEM was well defined	1.21	N.S.	N.S.
Specificity of INTENT for use of information	N.S.	N.S.	N.S.
Estimate of existing PUBLIC KNOWLEDGE was high	1.67	1.87	N.S.
Degree of INTERNAL KNOWLEDGE	N.S.	N.S.	N.S.
User Constraints on Searches			
The LANGUAGE of answers was restricted to English	1.59	1.75	N.S.
TIME LIMIT was placed on years searched	1.41	2.0	N.S.
Question Characteristics			
For questions judged to be:			
of low semantic CLARITY	1.27	N.S.	N.S.
of low syntactic CLARITY	1.35	N.S.	N.S.
of low SPECIFICITY of query part	N.S.	N.S.	N.S.
of low SPECIFICITY of subject part	1.82	2.13	N.S.
of high COMPLEXITY indic. on a scale	1.71	2.27	N.S.
of high COMPLEXITY indic. by no. of concepts	1.93	2.16	N.S.
With many PRESUPPOSITIONS indic. on a scale	N.S.	N.S.	N.S.
With many PRESUPPOSITIONS indic. as to no.	1.45	N.S.	N.S.
Searchers Characteristics			
Frequency of DIALOG searching	N.S.	N.S.	N.S.
High scores on RAT, test of verbal association	1.60	N.S.	N.S.
Scores on SRT, test of symbolic reasoning	N.S.	N.S.	N.S.
Learning Style Investory (LSI)-preferences for the following learning style:			
Low for Concrete Experience (CE)	1.41	2.0	1.96
Reflective Observation (RO)	N.S.	N.S.	N.S.
High for Abstract Conceptualization (AC)	1.29	N.S.	3.27
Active Experimentation (AE)	N.S.	N.S.	N.S.
High for Abstractness over concreteness (AC-CE)	1.41	N.S.	2.47
Experimentation over reflection (AE-RO)	N.S.	N.S.	N.S.
Search Tactics and Efficiency			
No. of COMMANDS	N.S.	N.S.	N.S.
High no. of command CYCLES	1.18	N.S.	N.S.
Low no. of Search TERMS	1.28	N.S.	N.S.
Below average PREPARATION TIME	1.15	N.S.	N.S.

TABLE 32. *(Continued)*

VARIABLE When the following happened . . .	Relevance Odds	Precision Odds	Recall Odds
	Increased by a factor of		
ONLINE CONNECT TIME	N.S.	N.S.	N.S.
Below average TOTAL TIME used	1.23	N.S.	N.S.
Utility Measures (as assessed by users)			
Results WORTH more time than it took	1.42	2.40	N.S.
Took less TIME than average to evaluate	1.14	N.S.	1.64
DOLLAR VALUE for results was high	1.34	1.69	N.S.
High contribution to PROBLEM RESOLUTION	1.88	3.21	N.S.
High level of overall SATISFACTION	1.92	2.49	N.S.
Size of Output (per question)			
No. of RELEVANT items was high	2.95	4.43	N.S.
No. of PARTIALLY RELEVANT items was high	2.43	2.90	−37%
No. of NOT RELEVANT items was low	4.54	5.88	N.S.
Total no. of EVALUATED ITEMS was high	1.21	N.S.	N.S.
No. of items NOT EVALUATED was high	1.39	N.S.	−42%
No. of TOTAL ITEMS RETRIEVED was high	1.39	N.S.	−42%

lyzed, although they are almost surely present. Thus, it could not be correct to suppose that for two variables that have significant relevance odds, the increase in odds involving both could be obtained by multiplying the two odds; in other words, we cannot assume that increase in odds involving P(A) and P(B) are equal to P(A) × P(B).

Users

Users' Context. Relevance odds increased for questions where the users indicated that the problem was well defined. Both relevance and precision odds increased where users' estimate that information could be found in existing public knowledge was high. The degree of internal knowledge about the problem at hand made no difference. Recall odds were not significantly effected by any context characteristics.

For the well-defined problem relevance odds increased 21%, and for high estimate of public knowledge relevance odds increased 67% and precision odds 87%.

Question Constraints. Restriction on language of answers to English and restriction as to the years of publication increased relevance and precision odds. Recall odds were not significantly affected by either constraint.

When the language of answers was restricted to English relevance odds increased by 59% and precision odds by 75%, and when a time limit was placed on years searched relevance odds increased 28% and precision odds doubled.

User Applications. There were no statistically significant effect on precision for questions of different application as indicated by users. The mean overall precision for different applications was as follows: faculty research 50%, graduate study 49%, industry 49%, general 70%. Overall recall cannot be calculated, because we don't know what is left in the database, and relevance, precision and recall odds cannot be calculated, because applications have more than two classes.

Comparison between Users and Searchers. To a large degree searchers were able to assess the context of questions the same way as users. Searchers and users had a high degree of agreement on problem definition and intent, and a somewhat lower degree of agreement on existence of public knowledge; users estimated a higher probability that information could be found. As expected, in the degree of internal knowledge about the problem at hand, users scored much higher than searchers.

In comparison: 58% of users and about 50% of all searchers considered the given problem underlying a question as well defined; 45% of users and about 50% of all searchers scored the intent as "open to many avenues;" 60% of users and 58% of outside and 30% of project searchers considered that there was close to certainty that information requested could be found in public knowledge. For 45% of questions users considered themselves as quite knowledgeable about the problem at hand, while the outside searchers considered themselves on the opposite end, as having quite low internal knowledge in 60% of cases and the project searchers in 80% of cases.

Questions

Classification of Questions. A group of judges had a fair agreement (i.e., far from random) in characterizing the questions as to degree of clarity, specificity, complexity, and presuppositions present. (We cannot say if the agreement was "high" or "low," because of lack of a norm.) Searchers (as represented by classification judges) seem to be able to distinguish questions on the basis of these characteristics and classify the questions accordingly in a fairly uniform way. The classification scheme as postulated seems to be realistic.

While the variances were not large, the largest disparity was on the judgment on question specificity. Surprisingly, there was also a relatively large disparity on the judgment of how many search concepts were in a question. Judges seem to agree less on how specific a question is or how many search concepts it has than on how clear and complex it is or how many presuppositions it has. This may explain in part the low degree of agreement in selection of search terms.

Question Characteristics. Relevance odds increased in questions of low clarity, low specificity, high complexity, and many presuppositions. Of these, precision odds increased in questions of low specificity and high complexity. Recall odds were impervious to any question characteristics.

For questions of low semantic and syntactic clarity, respective relevance odds increased by 27% and 35%; for low specificity of subject part of the question relevance odds increased 82% and precision odds more than doubled (increased 2.13 times); for high complexity indicated on a scale (or by number of search concepts) respective relevance odds increased by 71% (or 93%) and precision odds increased 2.27 (or 2.16) times; for many presuppositions (measured by a number) relevance odds increased 45%.

Searchers

Cognitive characteristics. If the claims of standard tests are accepted, then searchers who scored high on word association had increased relevance odds. As to learning preferences, searchers who scored low on preferring concrete experience as a learning mode had increases on all three odds; searchers who scored high on preferring abstractness over concreteness in learning style had a large increase in relevance and recall odds. Ability of word association in searchers favored relevance odds and preference toward abstractness in thinking favored relevance and recall odds.

For searchers who scored high on Remote Associates Test, relevance odds increased by 60%. Scores on Symbolic Reasoning Test had no significant relation with any of the odds. For searchers who scored low on Concrete Experience in Learning Style Inventory, the respective relevance, precision, and recall odds increased by 41%, 100%, and 96% (this is the only case in the study when all three odds increased due to the same factor). When preferring Abstract Conceptualization, relevance odds increased by 29% and recall odds 3.27 times. High preference for abstractness over concreteness as a learning mode increased respective relevance and recall odds by 41% and 2.47 times.

Online frequency. Frequency of searching DIALOG had no effect on any of the odds, however, all the searchers were experienced, thus the comparison is not between experienced and inexperienced. Seventy-two percent of searchers searched DIALOG at least once a week, the rest twice a month or less.

Searches

Tactics and efficiency measures. As to tactics, higher number of cycles and lower number of search terms increased relevance odds; number of commands had no effect. As to efficiency, lower preparation and total time increased relevance odds, while online time had no effect. None of these had a significant effect on either precision or recall odds.

A high number of cycles increased relevance odds by 18% and a low number of search terms increased them by 28%. Below average preparation and total time increased respective relevance odds by 15% and 23%.

Different Search Types. Searches based on different sources (so called project searches) produced a significant difference in recall and no significant difference in precision. The best recall was from a search type done on the basis of a taped problem and intent statement by the users. The poorest recall was achieved when words from written questions were used as search terms without any elaboration (as if they were picked automatically).

The four types of project searches performed as follows: (i) searches based on the taped problem and intent statement had 32% mean recall and 63% mean precision; (ii) searches based on the taped statement plus written question had 23% recall and 63% precision; (iii) searches based on terms from written question with no elaboration had 18% recall and 57% precision; and (iv) searches based on written question plus thesaural elaboration had 25% recall and 61% precision.

Overlap

Overlap in search terms and items retrieved. The overlap or degree of agreement in selection of search terms by different searchers searching the same written question was relatively low. The overlap in retrieved items was even lower. Searchers tended to use substantial sets of different search terms in searches from the same question and retrieve even a more substantially different set of items as answers. In other words, different searchers seem to extract different language from a question (or see differing things in a question) and retrieve different sets from the same file searched. Only a very small percentage of the variation in the overlap of retrieved items could be attributed to the overlap in search terms.

The mean overlap on search terms was 27%, but the distribution was skewed toward the low end: in 44% of comparisons between searches the overlap was between 0% and 20%. The mean overlap in retrieved items was 17% and the distribution was even more skewed: in 59% of comparison between retrieved sets the overlap was between 0% and 5%. The mean overlap in retrieval of relevant or partially relevant items (disregarding the not relevant) was 18%, in 59% of cases it was between 0% and 5%. Only 2.5% of variation of overlap in retrieved items is explained by overlap in search terms.

Multiple retrievals. When the outputs of different searches for the same question are compared to each other, we found that most of the items are retrieved only once. However, the more often the same item was retrieved (by different searches for the same question), the more likely it was to be relevant. In other words, when different searchers searched the same question, the sets they retrieved had a low overlap or degree of agreement, however, for the items retrieved in common (i.e., for which there was multiple retrieval) the odds that they were relevant increased most significantly.

In considering retrieval from 9 different searches for a question, about 69% of retrieved items were retrieved only once. The odds that an item be relevant or partially relevant as opposed to not relevant when retrieved any number of times (i.e., 1 to 9) were about even, or 10 to 10. For items retrieved only once out of 9 searches, the odds were about 8 to 10. For items retrieved twice (out of 9), they were 15 to 10, for those retrieved three times, 14 to 10, and 4 times, 19 to 10. For items retrieved 5 times, the relevance odds jumped to 57 to 10, for 6 times, 44 to 10, and for 7, 8, or 9 times, to 57 to 10. In other words, an item retrieved 5 or more times (out of 9 searches) was over six times as likely to be relevant as an item retrieved once and about five times as an item retrieved any number of times. Similar differences in magnitude of odds were found in separate comparisons of multiple retrievals by 5 outside and 4 project searches.

Effectiveness Measures

Precision and recall. The mean values for search precision (57%) and recall (22%) in this study are similar to values found in many other studies. When precision and recall were plotted against each other, the two were not inversely related. To the contrary, when either precision or recall was considered as an independent variable, the other had a small positive correlation—the correlation between the two was 16%. As precision rose so did recall, but rather slightly, or as recall rose so did precision, also slightly. In our searches, those with higher precision tended to have higher recall and vice versa.

A low percentage of the variation in precision and recall could be explained by the variables used in this study through the application of regression analysis. The most important explanatory variables for precision were users' estimate of public knowledge, explaining 10% of observed variance, and searcher characteristic measured by Remote Associates Test, i.e., word association, explaining 5%. No other variable was significant. For recall only, one variable was mildly significant: the score on Learning Style Inventory, where searchers preferred abstractness over concreteness as their learning style; it explained somewhat less than 5% of the variance.

Utility measures. All five utility measures had a significant dependence on relevance odds, four on precision odds and one on recall odds. When relevance and precision odds increased, the users considered that: the results were worth more time than it took; the dollar values of results was high; the contribution to the problem resolution was high; and they had a high overall level of satisfaction. When relevance and recall odds increased, the users took less time than average to evaluate the results. High scores on four out of five utility measures are connected with increased relevance and precision odds, indicating a close positive connection between measures of relevance and precision on one hand and utility on the other hand.

For questions where users considered results worth more time than it took, relevance odds were higher 42% and precision odds 2.4 times; where the dollar value was high relevance odds were higher 34% and precision odds 69%; where the contribution to problem resolution was high relevance odds were higher 88% and precision odds 3.2 times; and where the level of overall satisfaction was high, relevance odds were higher 92% and precision odds 2.5 times. Where it took less time than average to evaluate the answers, relevance odds were higher 14% and recall odds 64%.

Size of Output. As expected, relevance and precision odds increased for questions with a higher than average number of relevant and partially relevant items and lower than average not relevant items. For questions where the total number of retrieved items was high, relevance odds increased, precision odds were not effected and recall odds decreased.

For questions with a high number of relevant items, relevance odds increased 2.95 times and precision odds 4.43 times; a high number of partially relevant items, relevance odds increased 2.43 and precision odds 2.9 times, while recall odds decreased 37%; a low number of not relevant items, relevance and precision odds increased respectively 4.54 and 5.88 times; a high number of evaluated items ($R + pR + N$), relevance odds increased 21%; a high number of not evaluated items (i.e., in addition to 150 which were sent to users), relevance odds increased 39% and recall odds declined 42%.

Levels of analysis

The micro or item-wise level of analysis using relevance judgment of each retrieved item as the basis for analysis proved to be much more powerful than the macro or search-wise level using precision and recall of searches as the basis. Item-wise level of analysis showed many more significant relations than the search-wise level. The question of what the measures of precision and recall do show and can show (i.e., their utility in research) should be re-examined. Furthermore, the whole notion of recall, as used at present in particular, should be seriously questioned.

Methods of analysis

The methods of statistical analysis commonly used in social sciences, particularly regression analysis and analysis of variance, proved to be disappointing in uncovering significant relations. We had more success with methods used traditionally in biomedicine, among them tests for rates, proportions and odds. We urge other researchers in this field to re-examine the choice of statistical methods.

Explanations

In a scientific context, explanation involves postulating and then confirming or refuting underlying causes for given effects. Explanation is a description and an accompanying test of confirmation or refutation of factors that contribute to given effects. The urge to speculate about why we found whatever we may have found is quite irresistible. However, at this stage of research on the topic, such speculation, be it ours or anybody else's, is a hypothesis rather than an explanation. Any such hypotheses may turn out to be confirmed and thus correct, but nevertheless, speculations (whether ours or those of the reader) until scientifically confirmed or refuted are at best conjectures or hypotheses. As plausible as they may be, as the state of knowledge stands now, all such hypotheses require critical testing.

We have fought the temptation to speculate as to explanations in this series of articles, because we deliberately limited the presentation here to models, methods, and associated findings. However, we invite and encourage speculations and hypotheses as a prelude to research (in order to confirm or reject them) and to theory building. Without such research the "explanations," "principles" and "self-evident truths" in this field will remain no more than educated guesses and speculations. Without confirmation anybody's educated guess is as good as anybody else's.

This brings us to a larger point: a need for a (scientifically acceptable or refutable) theory of information seeking and retrieving or searching, appropriate for contemporary context and needs. To be realistic and of use, such a theory must have at least the following characteristics. First, it must be consistent with (or even better: a part of) a broader theory of human information behavior, a theory that is based on people's patterns, use, strategies, heuristics, ambiguities, etc., involving information. Use of the computer and/or computer logic as a model (such as adapted in some of the cognitive science and artificial intelligence research) for such theory building is simply inadequate. Second, such theory must incorporate the context and content of information in addition to syntax and logic. It cannot be based on context and content free "messages" or rules (such as assumed in information theory). The problem context and the language has to be incorporated. Third, while the theory should seek to explain human information seeking and searching in general, there should be room to incorporate the status of individual and the apparent differences of patterns in individual concept formation, clustering and de-clustering. Fourth, such theory must have a strong dynamic interactive orientation, rather than be based on a linear, unidimensional flow of information or messages (such as in information theory or sender-message-receiver communication theories). Fifth, the description of interactions must account not only for human–human interactions (be they direct or meditated by records), but also for human–machine (human–systems) interactions, and moreover it must account for both in a consistent way. Sixth, it must allow for dynamic changes over time in underlying information structures (e.g., knowledge bases) in humans and in society at large. Such dynamic change should be associated with information additions and deletions, novelty and obsolescence, patterning, clustering and de-clustering, learning and inferences, and the like.

The type of theory characterized above follows suggestions made by Gardner [16], based on his synthesis and critique of current theoretical thinking on cognitive processes in a number of disciplines. It also corresponds to suggestions by Winograd and Flores [17] related to shift in orientation for the design of hardware and software systems to correspond to human language, thought and action.

At present a lot of data in support of such a theory exist (but still not enough), some empirical laws have been formulated and tested, and even bits and pieces of the theory itself have been suggested [16, 17]. Plausible explanations about this or that aspect have followed. However, an achievement yet to come is a more comprehensive, synthesizing theory using all of these as and if appropriate. This study should be considered as contributing ammunition for such theorizing. By itself it does not offer explanations.

Models

In selection of given variables for testing, we explicitly suggested models that enumerated factors involved with users, questions, searchers, and searches. For the most part the suggested models tested well, that is, the elements suggested by the models had by and large a significant relation with retrieval outcome.[1]

Users. Two out of four elements suggested in the model of user context had a definite significant relation with the outcome of the retrieval process. These were definition of the underlying problem, and estimation of existing public knowledge about the problem at hand. There was some indication that specificity of intent may also be of signifi-

[1] The basic difference between the notion of a theory and the notion of a model as treated here is that a theory explains and accounts, and a model merely enumerates the elements that interact to provide certain outcomes.

cance. Degree of internal knowledge showed no significant relation, i.e., for the moment it remained an unconfirmed element. This corroborates a number of related models of information seeking and information intentions whose prominent feature is a problem orientation.

Only two of the suggested constraints imposed by users on answers could be tested: restriction on language and limit on time of publication. Both had significant relations with the outcome.

Finally, as to users, variation in the intended applications proved to have no significant relation with outcome. However, most if not all, applications in this study were associated with some form of research. Thus we cannot say anything about different applications in general.

Questions. We postulated two aspects about questions, linguistic structure and a classification scheme of general attributes. As to the structure, we suggested that the questions in information retrieval consist of three parts: a lead-in, a query, and a subject. We did not report on the linguistic analysis in this series of articles, however, our preliminary conclusion is that the suggested linguistic structure did not work very well. For a number of written questions judges had difficulty in establishing which element of the question is the subject (i.e., main concept) and which is the object (i.e., a query about the subject).

As to classification, we postulated five attributes along which questions could be categorized: domain, clarity, specificity, complexity, and presupposition. Two things were tested: consistency of classification among a group of judges and the relation between questions in all categories, except domain, and the retrieval outcome. In general, the attributes were recognized with relatively high consistency among the judges, i.e., the scheme is realistic. The consistency was lowest in indication of specificity and number of concepts in a question; judges had most difficulty and largest disagreement in assessing how specific a question is and how many search concepts it has. This is related to the finding in low overlap in selection of search terms. Consistency was higher in other categories.

Each of the four categories tested had a significant relation with the retrieval outcome. In other words, different question types produced different retrieval results.

Searchers. Our variables, frequency of searching, Remote Associates Test, Symbolic Reasoning Test, and Learning Style Inventory model the four associated cognitive traits: online experience, word association, symbolic reasoning, and learning styles. Of the four, two (word association and learning style) had a significant relation with retrieval outcome; symbolic reasoning had an ambivalent relation and the degree of online experience was not significant, but all searchers were quite experienced. Both of these latter traits need further testing.

The test in which searchers' assessment of question context was compared with users', showed that searchers' can very well approximate users' judgment on problem definition and intent, and less well on estimate of public knowledge.

Overlap in selection of search terms and in retrieval of items by different searchers for the same question is modeled by the asymmetric measure selected to express the overlap. Significant differences among searchers were discovered: different searchers for the same question chose a substantially different set of search terms and retrieved a substantially different set of items.

Searches. The model of tactics and efficiency of searches is represented by the measures chosen. Of these, a number of cycles and of search terms (as to tactics), and preparation time and total search time (as to efficiency) had a significant relation with retrieval outcome. Number of commands had no relation and online connect time had an ambivalent relation.

As to basis for their construction, we modeled four types of searches based on: problem statement, problem statement plus written question, terms from written question without elaboration, and written question plus elaboration through thesaurus. There were significant differences between the four as to retrieval outcome.

General conclusions about models. While the models chosen in this study were by and large shown to be fertile, meaning that the chosen elements had a significant relation with the outcome of the whole retrieval process, we cannot claim that this is a complete model representing all or even the significant elements in information seeking and retrieving affecting the outcome. By necessity we had to limit our study to a certain number of elements. Glaringly missing in our study are at least two very important aspects and associated models: (i) pre-search interaction between users and intermediaries (e.g., interview) or between users and system interfaces, and (ii) rules associated with searching, be they heuristic rules for human searchers, deterministic or probabilistic rules for machine searching, or the interaction between the two.

Many more tests need to be done on models and elements studied here and on additional ones not studied here, before a comprehensive model of information seeking and retrieving could be built and confirmed. While numerous models on these processes are suggested in the literature and reviewed in Part I, the problem is that very few were actually tested and some were untestable to start with.

Research Agenda

A main component of the basic research agenda for information science for time to come should be (i) participation in development of a theory of information behavior as discussed above, and (ii) test of models of information seeking and retrieving involving the human elements, be they users or intermediaries. In turn, the success of the applied research agenda whose present aim is the design and evaluation of various interactive information systems or system components (e.g., distributed expert systems, intelligent interfaces) is predicated to a large, if not overwhelming degree, on the connection with and success of the basic research agenda as outlined in respect to theory and models.

To build a machine (including an intelligent interface with a machine) that does some information searching tasks at least as well as humans do, we must first study the patterns of human behavior, as well as the patterns that relate relevant texts (in whatever form or image) to questions and problems at hand. For instance, this may lead to understanding of what makes for a more effective search pattern and what for a less effective one. When the patterns can be formulated as rules, rule classes, or rule components, we must then understand the functional impact of these rules on the task at hand. The inherent fuzziness of such rules while increasing their value as an area of research, at the same time substantially increases the difficulty of the task. Nevertheless, such rules should be tested not only in a laboratory, but under realistic standards involving real-life circumstances, users, and systems. This research agenda corresponds closely to the one suggested by Belkin and Croft in their critical review of practice and research in retrieval techniques [18].

To encourage further research and a test of various hypotheses that can be formulated on the basis of our data, we are providing to those interested educational and research institutions a tape containing all of our data files, together with over 30 associated SPSSX programs.[2]

Practical Implications

There are some general implications of our results for the practice of searching for information, including for systems design of a searching interface.

The context of a question is confirmed to be important. This suggests that it is important for searchers (or interfaces) to explore the background of a question and get as much information as possible about the problem at hand and the intent in use of information. The user's estimate of how much information about all of this may be out there can be a contributing factor in intermediate or final evaluation. The search should be planned more on these context aspects than on the written question alone. There is more to a question than the written words expressing it. If a search is based only on that, it may be expected to do rather poorly.

Different categories of questions classified as to clarity, specificity, complexity, and presupposition of questions may be expected to have different performance levels. But searchers may have substantial disagreement among themselves as to how specific a question is, how many search concepts there are, and which ones should be selected as search terms. This suggests that there may be more than one "right" way to search a question and that consultation among searchers may be a wise approach in selection of what to search for. In such consultations, a large amount of disagreement may be expected.

Skills in word association and a preference for abstract thinking appeared to be important abilities in searchers with higher search performance. This suggests that cultivating semantic association, be it in the language in general or in a subject in particular, seems to be a profitable enterprise for searchers. Learning in terms of abstractions and generalizations also helps.

Cycles in searching showed a significant impact on outcome. This suggests that it may be quite important to view and review intermediate results as the search progresses and adjust the search accordingly. Restriction of language to English showed better results, as did restriction of time limit on publication data. Searches could be constructed accordingly with and without such restriction and reviewed as in cycles.

High marks on utility measures were generally connected with high relevance odds and high precision. This suggests that more extraneous materials in output of searches ("to catch everything") may not be a way to higher satisfaction assessment by users (in our sample, of course).

Finally, the disparity in what searchers were looking for and what they retrieved for the same question was quite large. Searchers tended to see different things in a question and find different answers, and when all were put together each searcher contributed to the totality of relevant answers. However, the odds that an item be relevant increased dramatically for items that were retrieved in common by several searchers for the same question. This suggests that multiple searching of the same question by different searchers or the same searcher but different strategies may be a valuable tactic, particularly if it then includes inverse ranking of output according to retrieved once, twice, three times and so on.

We are not sure at all if the searchers, the profession at large, or the system designers are fully cognizant of how little agreement there is in searching and how little overlap there is in a great many information processes over the same items (index terms, questions, searches, retrievals, citations, etc.). Since this low degree of overlap has been found in a number of studies, not only here, it may be quite worthwhile to explore it further, particularly for practical and design reasons.

Searching for information is far from being a science, and as yet the present heuristic rules or principles of searching as stated do not take into account some important aspects of what seems to be really going on. As yet, a plausible algorithm or even a reasonably comprehensive and consistent set of heuristic rules reflecting human information searching does not exist. Searching is still an art and a very imprecise art at that.

Acknowledgments

The complex and lengthy process of assembly of data for this project, including communication with users, administration of searchers and searching, and compilation of "raw" data was managed by Alice Y. Chamis and Donna Trivison. The programming for statistical analyses was done by Jun-Min Jeong, J. J. Lee, Moula Cherikh, and Altay Guvenir. Their contribution is fully and gratefully acknowledged.

[2] The magnetic tape in ASCII format can be purchased from Tantalus, Inc. Write to Paul Kantor, Tantalus Inc. 3257 Ormond Rd. Cleveland Hts., OH, 44118.

References

1. Saracevic, T., Kantor, P., Chamis, A. Y., Trivision, D. "A Study in Information Seeking and Retrieving. I. Background and Methodology," *Journal of the American Society for Information Science*. 39(3):161–175, 1988.
2. Saracevic, T., Kantor, P. "A Study in Information Seeking and Retrieving. II. Users, Questions and Effectiveness," *Journal of the American Society for Information Science*. 39(3):176–195.
3. Saracevic, T., Kantor, P., Chamis, A. Y., Trivison, D. *Experiments on the Cognitive Aspects of Information Seeking and Retrieving. Final Report for National Science Foundation Grant IST-8505411*. National Technical Information Service; (PB87-157699/AS). Educational Research Information Center (ED 281 530); 1987.
4. Bates, M. J. "Subject Access in Online Catalogs: a Design Model." *Journal of the American Society for Information Science*. 37(6): 357–376; 1986.
5. Zunde, P., Dexter, M. E. "Indexing Consistency and Quality," *American Documentation*. 20(5):259–267, 1969.
6. Markey, K., Demeyer, A. N. *Dewey Decimal Classification Online Project: Evaluation of a Library Schedule and Index Integrated into the Subject Searching Capabilities of an Online Catalog*. Dublin, OH: Online Computer Library Center (OCLC); 1986.
7. Dalrymple, P. W. *Retrieval by Reformulation in Two Library Catalogs: Toward a Cognitive Model of Searching Behavior*. Ph.D. Dissertation. Madison, WI: University of Wisconsin-Madison, 1987.
8. Katzer, J., McGill, M. J., Tessier, J. A., Frakes, W., DasGupta, P. "A Study of the Overlap Among Document Representations," *Information Technology: Research and Development*. 2:261–274; 1982.
9. Pao, M. L. "Comparing Retrievals by Kevwords and Citations," Proceedings of the Seventh National Online Meeting, Medford, N.J.: Learned Information: 1986; pp. 341–346.
10. Pao, M. L., Fu, T. T. W. "Titles Retrieved from Medline and from Citation Relationship," *Proceedings of the Annual Meeting of the American Society for Information Science*. 22:120–123; 1985.
11. Pao, M. L., Worthen, D. B. "Retrieval Effectiveness by Semantic and Citation Searching," *Journal of the American Society for Information Science*. (In press.)
12. Rapp, B. *A Comparison of Document Clusters Derived from Cocited References and Co-assigned Index Terms*, Ph.D. Dissertation. Philadelphia, PA: Drexel University, 1985.
13. McCain, K. W. "Descriptor and Citation Retrieval in the Medical Behavior Sciences Literature: Retrieval Overlaps and Novelty Distributions," *Journal of the American Society for Information Science*. (In press.)
14. Stepp, R. E., Michalski, R. S. "Conceptual Clustering of Structured Objects: A. Goal Oriented Approach," *Artificial Intelligence*. 28: 43–69; 1986.
15. Belkin, N. J., *et al.* "Distributed Expert-Based Information Systems: An Interdisciplinary Approach." *Information Processing and Management*. 23(5):395–409; 1987.
16. Gardner, H. *The Mind's New Science. A History of the Cognitive Revolution*. N.Y.: Basic Books; 1985.
17. Wineograd, T., Flores, F. *Understanding Computers and Cognition*. Norwood, N.J.: Ablex; 1986.
18. Belkin, N. J.; Croft, W. B.: "Retrieval Techniques." Ch. 4. in Williams, M., Ed. *Annual Review of Information Science and Technology*. Amsterdam: Elsevier; 1987. 109–145.