
Users' Information Needs At Different Stages of A Research Project: A Cognitive View

Peiling Wang

ABSTRACT

This paper reports on part of the results of a longitudinal study on document selection and use by faculty and graduate students during a research project. The study reveals that users only consulted half of the selected documents and cited one-quarter of the read documents; half of the users did not cite any. A further analysis of the users' information need statements provided some understanding of the changes in their cognitive structures across different stages of the projects.

INTRODUCTION

This study was carried out in a 2.5-year span to examine real users' document use during a research project. The purpose of this longitudinal study is to examine real users' cognitive behavior and address two issues: (1) how they select documents based on bibliographic information and (2) how they actually use the selected documents in subsequent stages. It is an exploratory study to identify decision criteria and rules in selecting and using documents in a task-evoked information retrieval (IR). Results on how user selected the retrieved documents and subsequently used these selected documents have been reported in a dissertation and three conference papers [Wang 1994; Wang and Soergel 1993; Wang and White 1995, 1996]; several journal articles are under preparation.

The participants in this study are academic users with real information needs related to their research projects. These users were working on typical tasks in an academic setting, such as journal articles, funding proposals, and dissertations. Although not all the users stuck to their original topics, 60% of them succeeded. As document use is concerned, about 20% of the retrieved documents were accepted; 50% of the selected were actually read; 25% of the read (13% of the selected) were cited in the final products; and about half of the users did not cite any selected document. Prompted by the phenomenon, the researcher further examined these users' information need statements across the stages of their projects. The purpose of this analysis is two-fold: (1) to explore

how user's knowledge structures change and (2) to explore how IR systems should incorporate these changes.

INFORMATION NEEDS AND KNOWLEDGE STRUCTURE

The term information seeking focuses on active seekers. When a user recognizes an information gap in performing a task, she is likely to seek information to bridge the gap in order to move to a desired state (Dervin 1980). In interacting with an IR system, the user is asked to specify her information needs. The dilemma, as Belkin, et al. (1982) point out, is that the user is unable to explicitly express her information needs because she is in an "anomalous state of knowledge" (ASK). Brookes (1980) proposes that a new modified state of knowledge can be achieved by the effect of information, well-known as:

$$K(S) + \Delta I = K(S + \Delta S)$$

where

$K(S)$ is the knowledge structure

ΔI is the increment of information

ΔS is the effect of the modification

His theory assumes that knowledge is a structure of concepts linked by their relations and information can be assimilated into this structure.

In IR, an understanding of user's $K(S)$ can optimize the design of the systems. From the cognitive viewpoint, the information-processing model holds that knowledge is organized in human's long-term memory (LTM) as a semantic network consisting of nodes (concepts) and links (relationships); the network can be accessed through activation (Best 1995). Concepts are the basic units of the LTM; concepts in LTM are permanent. The relations among the concepts are categorical, hierarchical, and associative. Ingwersen (1992) uses map to illustrate $K(S)$ in LTM. A concept can have different maps simultaneously in the mind of an individual; concepts and conceptual relations on individuals' cognitive maps change over time and across situations.

When a user carries out a project on a topic from literature search to a written product, her knowledge on the topic is gradually evolving from an ASK to a coherent state of knowledge (CSK). An understanding of the change in $K(S)$ along this process has important implications for IR. Based on ASK, users can express well what they already know and what they expect to know in either oral or written format. In any case, the user must have a vocabulary about the topic available in her LTM at the time of verbalization. This verbalized vocabulary is worth scrutiny because it represents the $K(S)$ of the user.

METHODOLOGY

Participants.—Twenty-five faculty and graduate students in the Agricultural and Resource Economics, University of Maryland, participated in the first study in Summer of 1992, when they were searching for literature for their projects. (Hereafter, the 1992 study.) Fifteen of the twenty-five, who had completed or were near completion of their projects, participated in the follow-up study in Spring of 1995. (This follow-up study will be referred to as 1995 study.) This paper is based on the fifteen users who participated in both studies.

Research Design.—All users in the department were invited to participate in the 1992 study when the Dialog, Inc. provided free online access to their databases. A pre-search interview (audio-recorded and transcribed) was conducted to obtain users' information needs. The search results were printed out in full bibliographic records. The participant was asked to perform the document selection task and to verbalize her thoughts simultaneously; the researcher was onsite during this process and both had a copy of the bibliographic citations. This process was audio-recorded and transcribed. The reference interview and the document selection process were mostly done within a day or two.

In the 1995 study, the fifteen participants who had finished their projects (two near completion) were interviewed. The interview solicited information on: (1) how the final product fit into the original project topic; (2) how each of the selected documents was actually used (read or cited); and (3) why they cited certain documents not retrieved in the 1992 study (randomly selected from their citation lists; some went through all the cites). This post-project interview was also audio-recorded and transcribed.

Data.—The data were extracted from the fifteen users' transcripts in the two studies. A user's need was presented first as a request during the reference interview; then, it was reiterated voluntarily during the document selection process; finally, it was mentioned frequently while justifying the use and nonuse of the documents during the post-project interview. For each participant, there are three sets of information need statements corresponding to the three stages: (1) request; (2) document selection; and (3) post-project. From these statements, an active vocabulary (V) corresponding to each set is identified: $V1$, $V2$, and $V3$ (see Figure 1).

Data Analysis.—The data analysis focused on users' active vocabularies across the three stages. The basic units of analysis are terms. A term is a verbal representation of a concept. It consists of a single word or a phrase (multiple-words). A term also represents a pre-coordinated concept. The first step is to extract the terms (the italicized words in figure 1). The second step is to compare $V1$ with $V2$ and $V1 + V2$ with $V3$ to identify total terms, new terms, dropped terms, and relations of the new terms to terms in previous stages. The term relationships considered are listed in figure 2. They are commonly seen in a thesaurus. A term and its different wording are considered as one term.

FIGURE 1

Excerpts of need statements of a user

Request: "How *regulations* affect *investment* in *pollution control*?" "Dynamic regulation, pollution reduction..." (V₁)

Selection: "There are two things: the *investment* and *regulatory choice* or *regulatory response*; the combination is what exactly the core of my model."

"... in terms of either *regulatory response* or *long term investment* issue—two things that are important to my topic."

"I am interested in *equilibrium*." (V₂)

Post Project: "The paper I was interested in was something truly *dynamic*. ... *lots of time periods*."

"One of the concepts I use in this paper is called *time-consistency* and this is one of the first articles ..." (V₃)

FIGURE 2

Term relationships

Synonym (ST): term *B* is interchangeable with term *A*
Broad (BT): term *B* is broader (in hierarchy) than term *A*
Narrow (NT): term *B* is narrower than term *A*
Related (RT): term *B* is associated to term *A*

Given the nature of the data-natural language expressions in ordinary narrative style, identifying unitary terms and their relationships was very difficult. And these difficulties are not encountered only by this study (Bates et al. 1993, 9). The rule of thumb has been to consider how a concept was experienced by the user and how it stands in the field. Thus, the Library of Congress Subject Headings (LCSH) – a pre-coordinate index system, several prestigious economics encyclopedias and dictionaries, as well as The Journal of Economic Literature Classification System (JELCS) were consulted in addition to looking at the terms within the context of the statements in making decisions.

To map vocabulary changes in a cognitive structure, it is necessary to apply semantic factoring for at least two reasons: (1) a user's vocabulary often consists of different facets at different stages; (2) some compound concepts are produced by introducing a new facet. For instance, a user requested literature on economics of biodiversity, which has two broader subject facets: economics and biology. Later, he decided to focus on *Latin American countries*. Thus, a new facet: country was added to his cognitive structure for the topic and a new map was pulled down. The narrower concepts that are not from genus-species, whole-part, or type-token relations are the product of combination of different facets. Most related terms are concepts from another facet.

Examples:

regulation

NT *dynamic regulation*

Facets for NT = policy : time

soil erosion

RT *pesticide* use (as a cause)

Facets for both terms = soil : agricultural chemicals

As a typical example of dropping facets or sub-facets, a user started with his topic on *income inequality and poverty measurement*. Both concepts were kept during the document selection and reading. At the end of the project when a journal paper had been submitted for publication, he restated his topic focus as follows:

I wasn't really aware that there was a difference between the two measures. ... It wasn't until I started looking at things fairly carefully. ... The basic difference is that *poverty measure* defines an absolute threshold level. ... We end up not saying much about the *poverty measure*.

After consulting the LCSH and the JELCS, it is confirmed that poverty and income inequality can belong to different facets.

RESULTS

The Users.—Among the fifteen users included in this paper, there are 8 professors, 6 doctoral students and 1 masters student (see note in Table 1). Their final products are 3 journal articles, 1 monograph, 1 book chapter, 3 technical reports, 1 grant proposal, 2 dissertations, 3 dissertation proposals, and 1 thesis. Most participants claimed that their final products fit well or close into their request statements after they reread the reference interview transcripts. Only four users explicitly indicated some changes in their topics: User 8 dropped a facet; User 20 found a niche within the previous chosen broad topic to fit data; User 21 expanded topic by bringing new facets; and User 25 moved topic to a broader level to fit data.

The Terms.—There are 471 unique terms extracted from the transcripts of the fifteen participants across the three transcripts. The size of an individual's vocabulary varies from 2 to 28 with an average of 14 in all three stages.

Distribution of Vocabulary.—At the level of counting terms in each vocabulary for each participant, overall, V1 (average 11) is comparatively smaller than V2 (16) and V3 (17). There is not a single pattern for vocabulary change in size. In each active vocabulary across the three stages, there are some changes: terms dropped (average 6 for V2; 15 for V3) and new terms introduced (average 11 for V2; 10 for V3). It is evident that both the size and content of the vocabulary are not static. (see Table 1)

Term Relationships.—New terms in each vocabulary are compared with the terms in previous vocabulary. Most changes are along NT and RT. All users introduced narrower terms (average 5 for V2; 4 for V3); most users enlarged their vocabularies with related terms (similar to NT). Comparatively fewer synonyms or broader terms were brought in. (see Table 2)

Facet Analysis.—The facet analysis is done on an individual's basis, that is, to identify the facets within one user's vocabulary rather than to find a standard set of facets to be applied to all users or the discipline or field. This is because the latter would require the type of efforts needed for a thesaurus developer and this study aims at the change of individual user's K(S) related to her task topic not to the field or discipline.

In general, the number of facets for a topic ranges from 2 to 9. The results showed less changes in facet compared with those in term. In V2 more facets were added than dropped while in V3 dropping and adding were close. What did not show up in the table is that all the users kept their main topic facets in their requests. (Table 3)

TABLE 1
Vocabulary changes

User ID	V ₁			V ₂			V ₃		
	T	-	+	T	-	+	T	-	+
2	17	6	8	19	18	1	8		
3	7	4	15	18	13	10	19		
4	26	20	20	26	27	7	26		
6	4	1	9	12	12	7	8		
7	9	3	10	16	5	5	19		
8	10	3	13	20	20	8	11		
10	14	9	8	13	14	20	28		
11	8	4	9	13	13	8	12		
12	12	8	12	16	18	11	17		
13	6	3	9	12	11	11	15		
16	14	5	15	24	16	3	16		
17	2	1	10	11	11	9	10		
20	14	8	8	14	18	8	12		
21	12	10	13	15	17	19	27		
25	11	9	2	4	11	17	19		
Σ	166	94	161	233	224	144	247		
Aver.	11	6	11	16	15	10	17		

Notes:

User ID is assigned according to professional rank; 2-11 are professors, 12-21 doctoral students, and 25 Masters student.

To decide total (T), dropped (-), and new (+) terms in each set, V₂ is compared with V₁; V₃ is compared with the combination of V₁ and V₂. Thus:
 $V_{2,T} = V_{1,T} - V_{2,-} + V_{2,+}$; $V_{3,T} = V_{1,T} + V_{2,+} - V_{3,-} + V_{3,+}$

TABLE 2
Relationships of new terms

User ID	V ₂				V ₃			
	ST	BT	NT	RT	ST	BT	NT	RT
2			1	7				1
3	1		12	2				7
4	3		10	7	1			4
6	1	1	3	4		3		2
7	1		7	2	1			1
8	5	1	3	4	2	1		3
10		2	1	5	1	1	12	6
11	3	1	3	2			1	4
12		1	1	10	1	1		1
13	1		4	4	3			5
16	4		3	8				3
17			6	4	1			4
20			3	5	1			2
21	1	1	7	4		3		9
25			1	1			2	8
Σ	20	9	71	61	11	12	66	55

Notes:

For users' academic ranks, see note in Table 1.

New terms in V₂ are compared with terms in V₁; New terms in V₃ are compared with terms in both V₁ and V₂.

TABLE 3
Facet changes

User ID	V ₁			V ₂			V ₃		
	F	-	+	F	-	+	F	-	+
2	4			4					4
3	7	1		6					6
4	5		2	7					7
6	4		2	6	3	1			4
7	4		1	5					5
8	3		1	4	2				2
10	6	2	2	6			1		9
11	4		1	5					5
12	3		2	5	2	1			4
13	3		4	7	2	3			8
16	4		2	6	1				5
17	2		2	4	2	2			4
20	4	2	3	5	3	1			5
21	5	1	1	5	1	1			6
25	4			4			4		8
Σ	42	6	23	79	16	14			82

Notes:

For users' academic ranks, see note in Table 1.

Terms in each vocabulary are grouped into facets; total facets (F), dropped facets (-), and new facets (+). The calculation:

$$F_2 = F_1 - F_{2,-} + F_{2,+}; F_3 = F_2 + F_{3,-} - F_{3,+} + F_{3,+}$$

Term Features.—In V₁, most terms are at a basic-level. That is, these terms from users' requests are somewhere in the middle of a conceptual hierarchy. Broader and narrower terms were introduced in subsequent stages.

Examples:

environmental degradation (V₁) →

NT soil erosion; ground water pollution (V₂);

coalition (V₁) →

NT exogenous coalition formation;

endogenous coalition formation (V₂) →

BT game theory; (broader than *coalition*)

NT payoff division; Shapley value (V₃).

The second interesting phenomenon: although most terms are topic-specific, quite often the discipline names or field/course-related terms were used to describe their needs. The following are such terms: economics, environmental economics, political science, biology, ecology, agronomy, hydrology, microeconomics, econometrics, equilibrium, game theory, and programming.

Examples:

I don't want biodiversity alone; I don't think you can find many on *economics* of biodiversity.

I was more on *agronomy* side then, but now I am more on *economics* side.

The third feature is that these users often emphasized on a specific document orientation, in their own terms: theoretical, empirical, methodological, data, research design, case studies, policy oriented, real world perspective, historical sweep, general principles, survey, qualitative analysis, and estimation. This preference in document orientation is situational and subject to change. As one participant talked about an unused selected document in the post-project interview,

now I am doing some case studies of regulation. So, that makes it [a case study dissertation] more important and more worthwhile.

DISCUSSIONS AND CONCLUSIONS

In-depth analysis of users' active vocabularies related to their information needs (V₁, V₂, and V₃) has come up with some interesting results. Before the discussions and interpretations go any further, some limitations need to be pointed out. First, it is inherited in qualitative research that verbal data collected in natural settings are rich, detailed and insightful, meanwhile, unsystematic and incomplete. Second, the participants in this study are from a relatively homogeneous user group. Bearing in mind the nature of the data, the results cannot be generalized; rather they should be taken as exploratory, suggestive, case-based and subject to further studies.

Size, Depth and Breath of K(S).—The active vocabulary should be viewed as a portion of the K(S) that is activated and in user's focal attention, because "only information in focal attention can be verbalized" (Ericsson and Simon 1993, 90). Based on the results presented above, it can be inferred that the actual vocabulary (should include the portion not activated) in each later stage is substantially larger in size than the previous one, broader and deeper in hierarchy, and wider in breadth. On the other hand, no matter how dynamic a user's vocabulary may be across stages during a project that is to go on, the main facets of the topic are stable. It seems that facets for a topic are less dynamic despite the contents of the facets are dynamic.

Factors Account for Changes in Vocabulary.—The following factors might have contributed to the phenomenon: V₁ < V₂ < V₃. First, by definition, K(S) at ASK is vague and limited in contrast to that at CSK. V₁, V₂, and V₃ were recorded along the spectrum of ASK to CSK. The changes in vocabulary reflect the change in K(S). Second, the users were in different memory retrieval

conditions. In V1, the users must be able to *recall* all the terms relevant to the topic. In V2 and V3, the users were looking at bibliographic citations, which might have served as cues. The task is more like a *recognition*. Experimental evidence indicates that *recall* task requires more cognitive effort than *recognition* (Wessells, 1982). This may be the main factor for the differences between V1 and V2 because the time lag was short and changes in K(S) were limited. Third, the user might have modified her needs in later stages after she failed to find relevant literature for some aspects. This fits into Taylor's (1968) fourth level of question formation, the compromised need. Last but not least, as Brookes' theory indicates, the information obtained through reading bibliographic citations or actual articles made a change in user's K(S).

Basic-level Vocabulary, Discipline Names, and Document Orientation.—The evidence in this study supports basic-level effect and indicates: (1) the user believes that the basic-level terms should be used in query formulation based on her mental model of the IR system; (2) the K(S) is build upon this basic-level vocabulary and expanded from that base. The basic-level vocabulary, studied extensively by Eleanor Rosch and others, is basic in perception, function, communication, and knowledge organization (Lakoff 1987, 47).

For the users who were working on topics of multidisciplinary interests, the need to filter out documents from other disciplines is obvious. To express this need, end users, without knowing much about IR mechanisms, are likely to use these discipline names in their searches. As Bates et al. reported, 23 percent of all subject queries by humanities scholars in the Getty Project contained discipline terms (1993, 21).

Document orientation is also a conscious need for many users depend on their tasks at hand. Expressed as part of the need statements as non-topic terms, users desired certain document orientations, such as theoretical, empirical, case studies, etc. The databases used in this study lack document orientation indexing (current status) to support effective searches at this level. As a result, the second most-used decision criterion for this user group was document orientation.

Implications.—The results in this study cannot be used directly in building intelligent IR systems but raise a question: how can IR systems, stored with potential useful information, help users in moving from ASK to CSK? User's needs along this spectrum vary and are difficult to predict. The focus should be on vocabulary development. Thus, a modification to current search process is proposed.

The IR interaction with users at ASK should start with obtaining the user's V1 (as usual) and be followed by presenting to the user with a new vocabulary (VIR) based on documents (not as usual). This VIR should have more contents (beyond basic-level terms) with different structures: hierarchy, semantic network, and facets-whichever is closer to user's K(S). With the help of VIR, the user can generate V2, which is a step further from original ASK and better than

V1 for effective search. This process can go on repeatedly. This idea is along the same line as the interactive thesaurus (Jones et al., 1995) and the searchers' thesaurus (Johnson and Cochrane, 1995), which intend to help users in term selection at the outset of a search with an online graphic thesaurus. This paper suggests to display a vocabulary relevant to the user's topic based on documents (not merely thesaurus) to help the user with K(S) development. If the search is not modified by V2 immediately, at least the hits can be ordered accordingly: with the most likely selected documents on top; or grouped by facet. Soliciting user's weight on facets may be more effective than term weighting search.

In addition, some improvement in indexing must be made to deal with the need for searching discipline or field names as well as document orientations. Both are much needed by end users but missed by IR designers. As important as improving search function, IR systems also should focus on helping users to build correct mental models and to create an interactive learning environment. For example, the system should be able to explain the terms used in a search: whether they are useful or useless in the search; and why.

Further Studies.—The proposed IR interaction suggested a VIR based on documents but not yet a way to arrive at this VIR, nor the display of this VIR. Therefore, further research is needed to investigate these two issues. Also, further studies on users' vocabulary development during a project should collect data more often to record a series of V_i ($i = 1, \dots, n$). A different method should also be used to investigate users' actual K(S).

ACKNOWLEDGEMENTS

Without my participants, this research would not have been possible. My attendance in the ISIC 96 is supported by the School of Information Sciences and the Office of Research at The University of Tennessee Knoxville. I also wish to thank the conference organizers for providing this research forum.

BIBLIOGRAPHY

- Bates, M. J., Wilde, D. N. & Siegfried, S. (1993), An analysis of search terminology used by humanities scholars: The Getty online searching project number 1. *The Library Quarterly*, 63(1), 1–39.
- Belkin, N. J., Oddy, R. N. & Brooks, H. M. (1982), Anomalous states of knowledge as a basis for information retrieval. *The Journal of Documentation*, 38(2), 61–71.
- Best, J. B. (1995), *Cognitive psychology* (4th. ed.). St. Paul: West Publishing Co.
- Brookes, B. C. (1980), The foundations of information science. Part 1. Philosophical aspects. *Journal of Information Science*, 2(3/4), 125–133.
- Dervin, B. (1980), Communication gaps and inequities: Moving toward a reconceptualization. In B. Dervin & M. J. Voigt (Editors), *Progress in Communication Sciences* Vol. II. Morwood, NJ: Ablex, 73–112.
- Ericsson, K. A. & Simon, H. A. (1993), *Protocol analysis: verbal reports as data*. Cambridge, MA: The MIT Press.
- Ingwersen, P. (1992), *Information retrieval interaction*. London: Taylor Graham.

- Johnson, E. H. & Cochrane, P. A. (1995), A hypertextual interface for a searcher's thesaurus. *Digital libraries '95 proceedings*, 77-86.
- Jones, S. & et al. (1995), Interactive thesaurus navigation: Intelligence rules OK? *Journal of the American Society for Information Science*, 46(1), 52-59.
- Lakoff, G. (1987), *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.
- Taylor, R. S. (1968), Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29, 178-194.
- Wang, P. (1994), *A cognitive model of document selection of real users of information retrieval systems*. Unpublished doctoral dissertation, University of Maryland, College Park, MD.
- Wang, P. & Soergel, D. (1993), Beyond topical relevance: Document selection behavior of real users of IR systems. *Proceedings of the 56th ASIS Annual Meeting*. NJ: Medford: Learned Information, Inc.
- Wang, P. & White, M. D. (1995), Document use during a research project: A longitudinal study. *Proceedings of the 58th ASIS Annual Meeting*. NJ: Medford: Learned Information, Inc.
- Wang, P. & White, M. D. (1996), A qualitative study of scholars' citation behavior. *Proceedings of the 59th ASIS Annual Meeting*. Forthcoming.
- Wessells, M. G. (1982), *Cognitive psychology*. New York: Harper & Row.