A Cognitive Study of Information Seeking Processes in the WWW: The Effects of Searcher's Knowledge and Experience

Hitomi SAITO and Kazuhisa MIWA
Graduate School of Human Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601 Japan
{hitomi,miwa}@cog.human.nagoya-u.ac.jp

Abstract

In this study, we investigated, through a cognitive psychological experiment and its protocol analysis, human cognitive processes of seeking information on the WWW and the effects of subject's knowledge and experience on the information seeking processes and performance. In our experiment, the subjects were divided into two groups: one comprising expert subjects and the other novice subjects. All of the subjects were given a general search task and a specific search task. In the experimental results, except for one exceptional subject, we could confirm significant differences between the two groups in the solution time, the number of pages searched, and the kinds of pages accessed. We also propose a behavioral schema for tracing a subject's searching processes. The behavioral schema consists of four behavior levels on the WWW: Search, Results-of-search, Page-following-results, and Page-following-pages. Each subject's behavior was described as a transition of nodes, each representing the subject's behavioral state, and six kinds of operators connecting two nodes: Search, Link, Return, Jump, Browse, and Next among the four behavior levels. The results of an analysis using the schema showed some distinctive subjects' behaviors such as a breadth-first search or a depth-first search. We also examined the descriptions the subjects' behaviors by the schema quantitatively by analyzing the transition rate from one node to another node at each behavior level. The results empirically suggested that a searcher's knowledge and experience do affect his/her information seeking behavior on the WWW.

1. Introduction

Ever since the first World Wide Web (colloquially "the WWW") browser Mosaic was released in 1993 [1], the WWW has been developing explosively. According to a report by Cyveillance, the number of Web pages on the Internet exceeded 2,100 million pages in July 2000, and is still increasing by over 7 million pages a day [2]. In order to use the vast information on the WWW efficiently, various tools such as search engines and mail magazines are available. In addition, new technologies continue to be developed such as Web data mining and searching technologies in search engines.

However, today's technologies and methods are not enough for users to seek information on the WWW efficiently. Part of the problem is that it is not easy to explain the information seeking behaviors on the WWW with the former technological framework alone since the WWW differs significantly from other information systems in terms of structure and features. Considering this, we regard the information seeking behaviors on the WWW as problem solving behaviors appearing in the process of discovering target information from a complex search space. By applying several frameworks of problem solving theories, we hope to examine the information seeking behaviors on the WWW in detail from cognitive viewpoints.

In this study, we examine the cognitive processes of the information seeking behaviors on the WWW, and the effects of a subject's knowledge and experience about these processes through a cognitive psychological experiment and its protocol analysis.

2. Background

Information seeking can be well understood as a problem solving behavior of searching through a data space and discovering target information. Ellis's work in 1989 is presumably the earliest research that considers information seeking as a kind of cognitive process [3]. Ellis interviewed academic social scientists to investigate information seeking patterns and identified six characteristics: starting, chaining, browsing, differentiating, monitoring, and extracting. Ellis implemented an experimental system based on the information seeking patterns he identified [4].

0-7695-1393-X/02 $17.00 © 2002 IEEE

After the study of Ellis, information seeking started to be examined by researchers and librarians in various fields. The effects of the searcher's knowledge and experience in information seeking also started to be examined. For example, Marchionini et al. in 1993 compared the information seeking behavior of search experts and the same behavior of domain experts, and explored the effects of domain and search expertise. They recorded the seeking behavior in hypertext or full-text CD-ROM databases by computer science, business/economics, and law. A qualitative analysis led them to some features about the information seeking behavior of search and domain experts. More specifically, the behavior of search experts was characterized as a problem-driven process. They expected the forms and locations of targets by utilizing system features. In contrast, the behavior of domain experts was characterized as a content-driven process. They sometimes used technical query terms based on their knowledge and expected possible answers. They discussed system designs based on the results of the experiments [5].

Sutcliffe et al. in 2000 conducted an experiment using the MEDLINE database, and examined effects of the knowledge of a search system on information seeking. They divided final year medical students into two groups based on their knowledge of the MEDLINE database and asked them to solve four search tasks. They then analyzed the experimental results and discussed differences between the two groups quantitatively. They also discussed implications of the results for the design of IR interfaces [6].

In addition to these studies on traditional systems, in recent years, the researchers have also focused on the information seeking on the WWW [7],[8]. As mentioned above, most research works on information seeking have mainly aimed at application to the design of systems and interfaces. However, only a few research works have explored its cognitive processes focusing on the information seeking itself. We believe that it is crucially important to analyze the information seeking process in detail, using the human problem solving theories that have been continuously developed in the communities of cognitive science and cognitive psychology.

In this paper, to examine some of the features of human information seeking, we propose a coding schema that describes a searcher's behavior on the WWW in a unified format. We also investigate effects of a searcher's knowledge and experience on information seeking on the WWW.

3. Experiments

3.1. Subjects

Twenty graduate students, comprising nine students majoring in cognitive science, six students in psychology, and five students in library information science, participated in the pre-test. The pre-test, as a preliminary survey, included three questionnaires about daily WWW usage, information seeking style on the WWW, and knowledge about search engines. We took ten students as subjects of the experiment to follow from the results of the pre-test. We assigned the five students with the highest scores and the five students with the lowest scores in the pre-test as Expert and Novice students, respectively.

3.2. Experimental Environment

The experiment was conducted in an environment capable of accessing the real WWW. As our WWW browser, Netscape Communicator 4.75 was used and as our search engine, ODIN was used [9]. When the subjects were asked about their experience in using ODIN, almost all of them answered that they had never used it. Therefore, we believe that the subjects' behavior di d not depend on specific characteristics of ODIN as a search engine. The experimental environment is shown in Figure 1. In the experiment, the subjects' verbal protocols, behavior, PC screen data, and URLs of browsed Web pages were recorded.

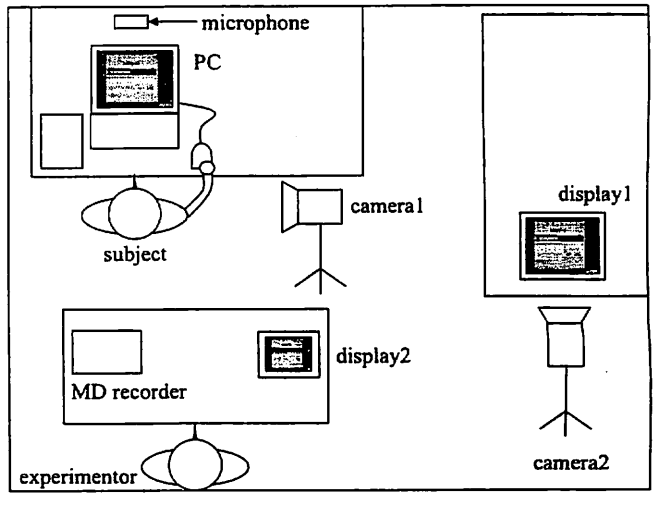

Figure 1. Experimental Environment

3.3. Tasks

All of the subjects were given the following two search tasks. One was a general task and the other task was a specific task.

- General task

  In a traditional South Korean wedding, seeds of vegetables are thrown toward the bridal couple. What are

seeds of plants? Also why the seeds were thrown?

- Specific task

  The ecology of a certain living thing, which has became clear from an incident in a foreign country, has given a shock to researchers in related fields. The living thing has a strong poison to affect human bodies. Although it is phytoplankton, it has different characteristics from the ordinary type of phytoplankton. The main feature of this living thing is to morph 24 times. Find out the formal name of this living thing. In addition, find out the literal meaning of the formal name, alias of the formal name, and name of the person who discovered the fearful ecology of the living thing.

We defined the general task as a task having a solution not requiring special knowledge. In contrast, we defined the specific task as a task having a solution requiring knowledge of a specific domain. We expected the processes of the Experts and Novices to differ on the basis of the characteristics of the tasks.

3.4. Procedure

After experimental instructions and training to verbalize the protocols, the subjects were asked to solve both search tasks using the WWW. Each subject was required to use ODIN as the search engine, to start from the search engine, and not to enter a URL directly. During the experiment, the subject was asked to verbalize his/her protocols. The subject was instructed to add pages about the target information to the bookmark menu. The time limit of each task was basically 20 minutes, but if the subject could not solve the task within the given time, an additional 20 minutes were given to the subject. After solving the two tasks, the subject answered a questionnaire. In the questionnaire, the subject was asked whether he/she had already known the answers of the given tasks prior to the experiment. The subject was also asked to report his/her strategies and methods employed while solving the tasks.

4. Results

4.1. Performance

The experimental results are shown in Table 1. In Table 1, the subjects are ranked according to their total scores in the pre-test. The "Result" column shows whether the target had been discovered or not. The "Time" column shows how many minutes were taken until the target was discovered. The "Number of pages" column shows how many pages were browsed, and the "Kinds of pages" column shows how many kinds of pages were browsed until the target was discovered. While counting the "Number of pages", we regarded the same pages that were searched more than one time as different pages; however, while counting the "kinds of pages", we regarded them as the same. If a subject could not discover the target, these columns show the total time and the total number of pages until the search was terminated.

In the general task, four Experts and two Novices discovered the target. In the specific task, only one Expert discovered the target. In each task, we compared the Experts' performance with the Novices' performance statistically based on the four indexes above: Result, Time, Number of pages, and Kinds of pages. Consequently, we found no significant differences between them.

Table 1. Experimantal Results

Subjects	Major	Results of pre-test				General Task				Specific Task			
		test1	test2	test3	Sum	Result	Time	Number of pages	Kinds of pages	Result	Time	Number of pages	Kinds of pages
E1	LIS	13	34	31	78	×	2400	184	106	×	2400	118	64
E2	LIS	10	34	31	75	○	720	41	33	○	1040	58	52
E3	LIS	12	37	23	72	○	1047	44	27	×	2400	99	54
E4	COG	13	26	29	68	○	1340	57	33	×	2400	187	87
E5	COG	16	25	26	67	○	2108	71	29	×	2400	83	62
N1	PSY	13	18	8	39	○	1287	109	72	×	2400	190	88
N2	PSY	6	20	12	38	○	2260	95	70	×	2400	106	56
N3	PSY	5	18	13	36	×	2400	118	66	×	2400	107	63
N4	COG	6	13	9	28	×	2400	104	68	×	2400	86	57
N5	COG	6	11	8	25	×	2400	99	60	×	2400	83	53

LIS:Library and Information Science

COG:Cognitive Science

PSY:Psychology

However, there were substantial differences between the basic search strategy of one subject E1 and the strategies of the other subjects. E1 used two or more browsers to seek the target whereas the other subjects used one browser. This difference is clearly seen in Table 1. E1 searched through a lot more pages than any of the other subjects. Therefore, E1's behavior possibly differed qualitatively from the other subjects' behavior. For these reasons, E1 was removed from the following analysis.

Again we compared the Experts' performance with the Novices' performance statistically removing the result of E1. As a result, in the general task, there were significant differences between the Experts and Novices in Time ($U = 2, p < .05$), in Number of pages ($U = 0, p < .05$), and in Kinds of pages ($U = 0, p < .05$). In the general task, the Experts discovered the target faster, and by referring to fewer pages, than the Novices. This difference was significant especially in the number of pages and the kinds of pages. The results shown above suggest that there are substantial differences in the searching behavior of Experts and Novices. Next, in order to clarify these differences in the behavioral processes of Experts and Novices, we describe the subject's searching behavior by introducing a behavior schema.

4.2. Behavior schema

4.2.1. Problem Solving Graph

In 4.1, we compare the Experts' and Novices' performance. As a result of the analysis excluding one exceptional subject, we can statistically identify significant differences between the Experts' and Novices' performance. Next, we investigate features of cognitive processes in information seeking.

In this study, we propose a behavior schema that describes cognitive processes in information seeking, and examine features of the seeking behavior on the WWW based on the schema. In human problem solving studies, there have been many schemas describing various problem solving behaviors. The Program Behavior Graph (PBG) proposed by Newell and Simon in 1972, is known as one of the most fundamental schema [10]. The schema proposed in this research is constructed based on the PBG. Figure 2 shows an example description of E2's searching behavior based on the behavior schema. Below, we explain important concepts used in the schema.

4.2.2. Four behavior levels

In this schema, each subject's behavior is described as a transition in four behavior levels: Search, Results-of-search, Page-following-results, and Page-following-pages. Each level corresponds to one of four fundamental behaviors in searching the WWW space by a search engine.

- Search: Searching by a search engine.
- Results-of-search: Browsing the results of the search.
- Page-following-results: Browsing a page selected from the results in the Results-of-search level.
- Page-following-pages: Browsing a page connected by a link with the page selected in the Page-following-results level.

4.2.3. Node

A node represents a subject's behavioral state. In our schema, each node corresponds to a state of a searcher's referring page. There are two kinds of nodes as shown in Figure 2. A new node is indicated with a square of bold lines ((1) as an example), whereas an overlapped node is indicated with a square of thin lines ((2) as an example). A new node is a node newly searched, and an overlapped node is a node repeatedly searched.

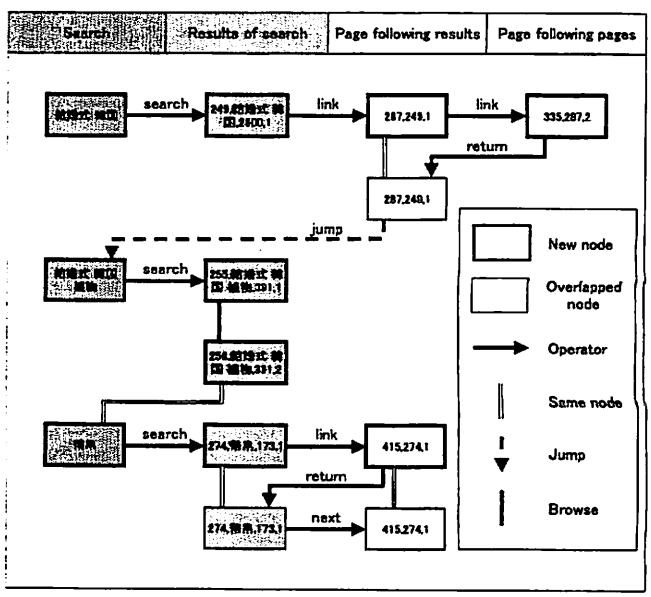

Figure 2. The behavior schema

4.2.4. Operator

Generally speaking, a subject's behavioral state is transferred by applying an operator. In our description, an operator is responsible for connecting two nodes. We define the following six operators.

- Search: Searching by a search engine

- Link: Going to a page connected with a link
- Next: Going forward to the next page after a subject had gone backward
- Return: Going backward to the last page that a subject had just visited
- Jump: Revisiting a page that a subject had visited previously
- Browse: Browsing search results that a subject had just obtained

Four of the operators, i.e., Search, Link, Next, and Return, represent movements between two different levels as shown in Figure 2. They are described with arrows of the same type. On the other hand, since Jump and Browse represent movements crossing two or more levels and within one level, respectively, they are described with different arrows from the other operators.

Nodes (1) and (2) in Figure 2 are the same nodes. These same nodes are connected with a double bar.

4.3. Qualitative analysis based on a behavior schema

As a result of describing all of the subjects' behavior based on the behavior schema proposed above, we identified a common behavioral pattern and individual characteristics that differ between Experts' and Novices' behavior.

4.3.1. Common behavioral pattern

The common behavioral pattern was observed across every behavior level of all of the subjects. We call this common behavioral pattern a basic unit. Figure 3 shows the basic unit.

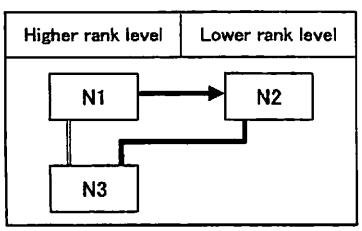

Figure 3. The basic unit

In Figure 3, a basic unit is defined as the following process: first transfer from N1 to N2, and then return from N2 to N3, which is the same node as N1. We can characterize the differences of Experts' and Novices' behavior by analyzing how global structures of subjects' behavioral processes are organized by combining these local basic units.

4.3.2. Common Experts' processes

The behavioral processes of Experts consist of sets of basic units that are organized systematically across several behavior levels. Figure 4 shows the processes of E2, E3, and E4, who are the top three subjects shown in Table 1 except for E1. In Figure 4, the smallest polygon shows a basic unit. A larger polygon shows a nested construction of basic units. As can be confirmed in Figure 4, the behavioral processes of Experts are constructed systematically by combining basic units. We can also see that the global structure of the behavior reflects a nested construction of basic units.

To explain why the processes of Experts can be characterized by a nested structure and to identify what kinds of cognitive factors underlie the nested structure, further analysis is necessary.

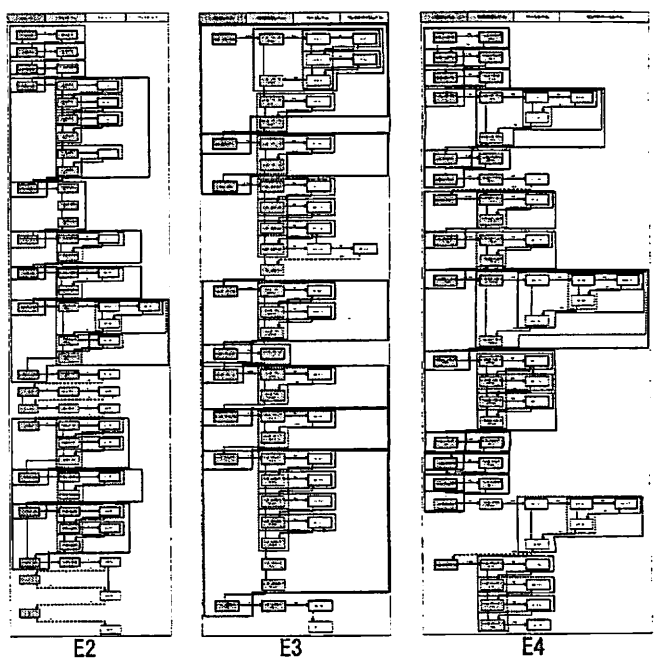

Figure 4. Three Experts' searching processes

4.3.3. Common Novices' processes

On the other hand, Novices' behavioral processes are substantially different from Experts' processes. We cannot confirm the systematic construction of processes seen in the Experts' behavior. As examples, we indicate two representative patterns observed in Novices' searching processes.

The first pattern can be understood as a depth-first search. Figure 5 shows an example behavior when a Novice

subject, N3, solves a general task. The horizontal movement in our schema represents the behavior of following links of pages. Accordingly, we can understand, in Figure 5, that subject N3 successively browses pages by following links, while fixing one specific page as a root node.

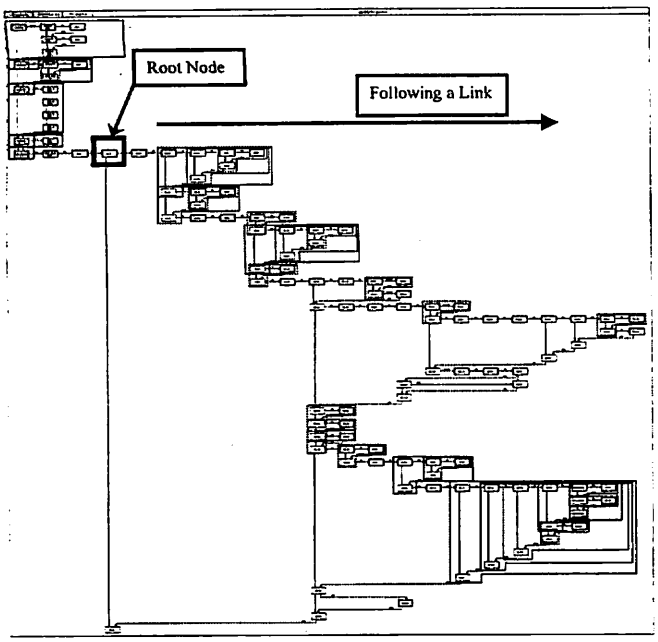

Figure 5. Novice N3's searching processes

The second pattern can be understood as a breadth-first search. Figure 6 shows an example behavior when a Novice subject, N4, solves a general task. The vertical movement in our schema represents the behavior of searching and browsing the results of the search. The point is that subjects N3 and N4 are unable to construct a similar systematic search pattern as observed in the Experts' processes above by tracing the search processes.

4.4. Quantitative analysis of transition patterns

As shown above, we found, through a qualitative analysis of subjects' behaviors using our behavior schema, that Experts' searching processes consist of well-organized sets of the basic units as shown in Figure 4. However, we also found that Novices' processes characterized as a depth-first search or a breadth-first search, do not consist of such systematic searching processes. Accordingly, to discuss the difference clarified above quantitatively, we calculated the ratio of transition from one node to another among the four behavior levels defined in 4.2.1.

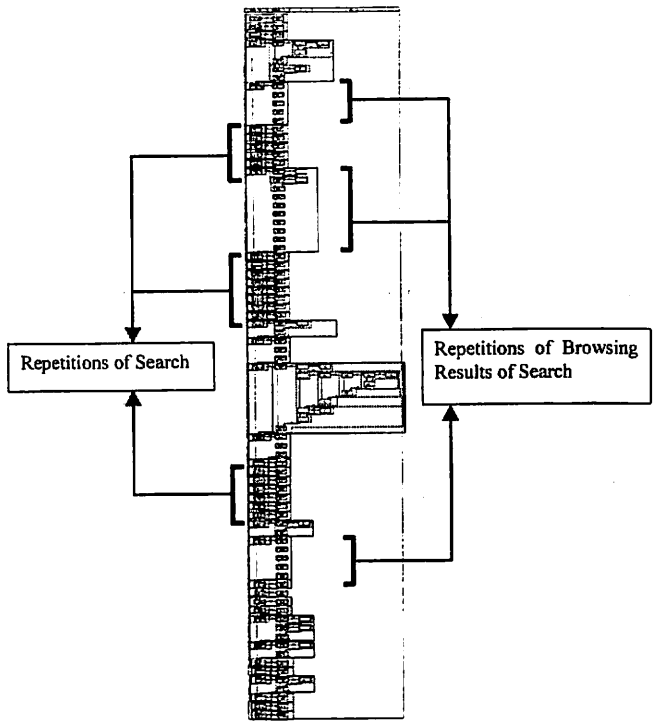

Figure 6. Novice N4's searching processes

Table 2 shows the transition ratios of the Experts except for E1; Table 3 shows those of the Novices. The bold and underlined cells indicate salient transition patterns. In Table 2, we understand that the major transition patterns, such as the transition from the Results-of-search level to the Page-following-results level and the transition from the Page-following-results level to the Results-of-search level, correspond to the movement constrained by the basic unit shown in Figure 4. On the other hand, in the Novices' cases shown in Table 3, we cannot confirm this kind of correspondence. That is, in the Results-of-search level, the transition ratios to three levels: the Search level, the Results-of-search level, and the Page-following-results level, are almost the same; moreover, in the Page-following-pages level, the transition ratio within the identical level is relatively high. When subjects control their behavior by a breadth-first search, the transition ratio is expected to increase from the Results-of-search level to the Search level and within the Results-of-search level. Additionally, when subjects control their behavior characterized as a depth-first search, the transition ratio increases within the Page-following-results level. Therefore, the tendencies shown in Table 3 provide us quantitative support about the qualitative nature of Novices' searching processes, as pointed out in 4.3.3. An analysis of the transition pattern of behavioral nodes confirms the results of the qualitative analysis using the behavior schema.

Table 2. The transition ratios of Experts except for E1

		transition from			
	Behavior levels	Search	Results of search	Page following results	Page following pages
transition to	Search	0	0.32	0.13	0
	Results of search	**1**	0.14	**0.61**	0.11
	Page following results	0	**0.55**	0.01	**0.58**
	Page following pages	0	0	0.25	0.28

Table 3. The transition ratios of Novices

		transition from			
	Behavior levels	Search	Results of search	Page following results	Page following pages
transition to	Search	0	**0.33**	0.01	0.02
	Results of search	**1**	**0.33**	**0.76**	0.03
	Page following results	0	**0.34**	0.05	0.29
	Page following pages	0	0	0.18	**0.66**

5. Conclusions

In this study, we carried out a protocol experiment to understand features of the cognitive processes underlying the information seeking on the WWW and effects of a searcher's knowledge and experience on the performance and processes. We conducted both quantitative and qualitative analyses on the experimental results. In the experimental results, by removing one exceptional subject, we could confirm significant differences in the solution time, and the number of pages and kinds of pages searched between Experts and Novices. Based on a confirmation of the differences in the final performance levels, we also tried to analyze the subjects' behavioral processes and identified qualitative and quantitative differences in the processes. The Experts organized their behavior by constructing a global structure systematically from local basic units, whereas the Novices did not. As our future work, we would like to make a more detailed description of the searching processes by analyzing our subjects' verbal protocols.

References

[1] Arms, Y. William. 2 The Internet and the World Wide Web. Digital Libraries. London, 2000, p.21-38, 344p.

[2] Cyveillance Press Releases. Internet Exceeds 2 Billion Pages, <http://www.cyveillance.com/web/us/newsroom/releases/2000/2000-07-10.html>,2000.

[3] Ingwersen, Peter. Information Retrieval Interaction. London, p.246, 1992.

[4] Ellis, David. A Behavioural Approach to Information Retrieval System Design. The Journal of Documentation. Vol.45, No.3, pp.171-212(1989).

[5] Marchionini, Gary, Dwiggins, Sandra, Katz, Andrew, Lin, Xia. Information Seeking in Full-Text End-User-Oriented Search Systems: The Roles of Domain and Search Expertise. LISR, Vol.15, No.1, pp.35-69(1993).

[6] Sutcliffe, A. G.;Ennis, M.;Watkinson, A. J.. Empirical Studies of End-User Information Searching. JASIS, Vol.51, No.13, pp.1221-1231(2000).

[7] ODIN Home Page, <http://odin.ingrid.org/>.

[8] Newell, Allen, Simon, A., Herbert. Human Problem Solving. New Delhi, p.920,1972.