Presented at the ASIS Annual Conference, 1999, Washington, DC
November 1, 1999

Factors Influencing Web Search Strategies[1]

by

Marilyn Domas White, Associate Professor
College of Library and Information Science
University of Maryland
College Park, MD 20742
E-mail: whitemd@wam.umd.edu

and

Mirja Iivonen, Professor
Department of Information Studies
University of Oulu
POB 1000
Fin-90401 Oulu, Finland
E-mail: mirja.iivonen@oulu.fi

[1]Both authors share equally in conceptualizing and writing this paper.

Introduction

The Internet has changed information searching considerably. If information searching in a traditional search environment, such as online databases and library catalogs, is still regarded mainly as a competency area of professional intermediaries, searching on the World Wide Web (web) is, at least theoretically, every man's right and commonplace. On one hand the web offers new possibilities for searching for information searching. It provides access to a huge amount of information and can be accessed (with the appropriate equipment) from everywhere and at any time. For example, people in Finland can easily follow what is going on in the United States in a wide range of subjects.

On the other hand, the web also creates an illusion about finding correct or appropriate information easily. New difficulties in information searching arise because of the enormous, fuzzy, and rapidly changing nature of the web. Users often become overloaded with too much information, too many sites, and too many links to be followed. They may experience difficulties in evaluating and selecting relevant documents. Many times they do not even know where or how to search. Users may also become frustrated because of badly designed web sites where relevant information is not to be found. At the same time that the Internet brings information close to the user, it also demands that he pay more attention to evaluating, filtering, and selecting documents.

On the basis of previous studies, major features of information searching on the web can be summarized briefly as follows:

1) *Known sites are important for users.*
Users go to the same sites from week to week (Tillotson et al. 1995) and use many landmarks for helping their search (Fidel et al.1999). Checking the web-site of the parent organization is a typical tactic for web searching and a unique search tactic of the web (Hsieh-Yee 1998). Tracing links from known and familiar sites is one of the main search strategies used on the web (Wang & Tenopir 1998, Wang et al. 1998).

2) *Users are browsing on the web.*
Users follow hyperlinks and use other heuristics (Catledge & Pitkow 1995). Users conduct swift and flexible searches and make quick decisions about where to click next (Fidel et al. 1999). Browsing the search result is a popular search tactic when too many pages are found (Hsieh-Yee 1998).

3) *Users use search engines often.*
Users often prefer to start their search by a search engine (Hsieh-Yee 1998). The use of search engine and modification of search statements is another main search strategy on the web (Wang & Tenopir 1998, Wang et al. 1998).

4) *Users prefer rather simple search statements and do not plan their searches*.
Users employ rather simple search statements on the web. They use only few search terms and lack motivation to modify complex search statements (Spink et al. 1998). When nothing would be found, they are ready to change one word to another (Hsieh-Yee 1998). The interactive nature of the web supports users' belief that there is no need to plan a search beforehand (Fidel et al. 1999).

5) *The success rate in web searches varies.*
The success of searches varies according to search questions (in Hsieh-Yee's study (1998) from 49 to 80 percent). Moreover, many times users are convinced that they have found the correct information

even when it is incorrect (Wang & Tenopir 1998, Wang et al. 1998).

This paper reports on a study of the influence of question-related variables on a web user's choice of search strategy in the initial stage of a search. The specific research questions are: 1) Do web users adapt their overall strategy in searching the web or do they use one approach habitually, regardless of the question? 2) What search strategies do web users prefer in the initial stages of a search? 3) Do web users differentiate their initial choice of search strategies based on the type of search question? 4) How is level of difficulty related to the question typology (and the question variables) used in this study?

In the web environment we suggest that at least two features of search questions impact on the chosen of the initial search strategy. The first feature is the openness of the question. In some cases the question is closed: the exact answer or answers exists and is wanted, and the searcher has little discretion in choosing alternatives. Often these are questions eliciting facts. In other cases, the question is open: there is no one exact answer, and the searcher has to develop an acceptable answer. There may be documents about the topic, and the user has to study them and combine information available through one or more sites. The second feature of search questions that impacts on the search strategy is the predictability of the source. In some cases, users knows with high probability of success where the answer can be found; in other cases they have to find a source to be able to find an answer. In these cases they cannot just go to certain web-sites, instead they have to use various search services.

Search question is defined as a verbal description of an information need. The *search strategy* is a plan or an approach to a search and can consist of several search tactics and moves. Search strategies can change during a search process. In this paper, the focus is on the participants= first decision in the initial stage of the search, referred to as the *initial strategy*. Three *initial strategies* were considered in this study: directly addressing the logical site, using a classified directory to identify a relevant site, and searching via a search engine to identify a relevant site. *Type of search question* refers to a combination of the features of the questions that are open or closed with predictable or unpredictable sources. This typology will be explained later. *Level of difficulty* in this study is derived by asking the participants to choose the five easiest and five most difficult questions; no independent measure is used and the variable is based solely on the participants= perception. Choice of initial search strategy is the dependent variable; the independent variable is type of question..

Search strategy always involves several steps. In the first stage, the searcher chooses one of three approaches to getting to a potentially relevant web site: he can type in an address and move directly to a likely site; he can search in a subject directory, such as Yahoo; or he can search in a search engine. Direct address usually takes only one step, but the searcher needs to know the likely content or linking structure of the site. He also needs to know a specific address or be able to create it based on his knowledge of the company or organization=s name, the address structure, categories, and abbreviations. Subject directories, such as Yahoo, are analogues to printed indexes or traditional library catalogues. Web documents or sites are classified and arranged into categories according to their topics. The directories are produced by human beings who select and organize documents, so the assurance is there that there is some quality associated with a site and some relationship to the topic of the category. Using a search directory is a multi-step process, but it involves recognition, not recall. A searcher needs to be able recognize topical or other categories that are likely matches for his question. Faced with the output of a search, as in direct address, he must deduce the content or linking structure of the site, often from cryptic search responses. Search engines provide access to massive indexes of all kind of information available on the web. Their indexes are based on programs that spider and surf on the web, continually seeking documents. The programs produce indexes automatically so there is room for mismatches between terms and content. Using a search

engine is another multi-step process, but requires the searcher to know the search features of the engine and to be able to translate his question into an acceptable search query. If he is discerning, he may also want to know the indexing policy of the search engine and the matching algorithm that is used. By emphasizing comprehensiveness, search engines forego evaluating sites. The selection between direct address, subject directory, and search engine often translates into a decision influenced by considerations about the amount of information or the number of documents to be searched, the degree of pre-screening, the probability of relevance, and the amount of effort involved in creating a viable search statement.

Methodology

Participants

The participants consisted of a total of 54 students in introductory courses in library and information science: 27 from an American university; 27 from a Finnish university. In each country the sample was adjusted to include only participants who indicated some degree of familiarity with the three types of overall web searching strategies addressed in this study. In the U.S., only students responding appropriately to a preliminary question about search strategy familiarity were randomly sampled; in Finland, all students in the core information access course completed the questionnaire. The Finnish sample was adjusted post hoc to exclude a participant who provided no data on this variable. Data were gathered in summer and fall 1998.

Based on responses to questions asking for demographic statistics, the participants were predominantly women (80 percent) with an average age of 30 (S.D. 8.6). They were experienced searchers, having spent about two and a half years searching (S.D. 1.4) and had generally learned by teaching themselves (69 percent) and/or by taking a class (30 percent). About 63 percent searched daily or several times a week, while an additional 30 percent searched weekly or several times a month. The percentage indicating a high degree of familiarity with the various strategies ranged from 48 percent for directories to 65 percent for search engines, to 76 percent for direct address.

Questionnaire

The questionnaire consisted of three parts. Part 1 asked participants to read 16 questions and, for each, to identify the search strategy they would try first for the question and their reasoning. For the same 16 questions, Part 2 asked the participants to identify the five easiest and five most difficult questions and to indicate their reasons for these choices. The participants did not have to rank the questions within each category. Part 3 asked questions related to personal characteristics of the participants, emphasizing their experiences in web searching. The questionnaire took approximately 25 to 35 minutes to complete.

The questionnaire was administered in two ways: some participants completed it in a supervised session; others completed it at home. In all cases, participants were advised to avoid collusion in their responses and to have their responses represent only their own judgment and reasoning.

The sample reference questions used in the questionnaire were devised by the researchers, based on their experience in searching the web and observing others- searching. They represent a combination of two factors: the closed/open nature of the questions, and the extent to which the source can be predicted (predictable, unpredictable). Sixteen questions were used, four from each type. Table 1 shows the questions grouped by category.

In addition, the questions were written to insure that neither national group had an advantage based on general cultural awareness of the subject matter of the question. In most cases, this was done

by using content that was assumed to be fairly universally known (See Question 9). In one case, the specific entity mentioned was changed to provide separate (but equal) American and Finnish versions. Question 16 asked the participant to identify the chief executive officer for Intel (American version) or Nokia (Finnish version). In each version the information requested is the same, but the company is a prominent company headquartered in the participant=s country.

Data analysis

The significance of the differences in the judgment of the difficulty of the questions and in the selection of search strategies was tested statistically. The testing method used was the chi-square test for k independent samples (See, for example, Siegel & Castellan 1988, pp. 190-200). The participants= explanations for choosing a particular strategy or deciding the particular level of difficulty of a reference question are included in this paper only to illustrate the reasoning behind the decisions. They will be analyzed in greater detail in another paper.

Results

Web Users= Adaptiveness

Some degree of familiarity with all three strategies was a prerequisite for the participants in this study. Familiarity or knowledge of the three strategies is one thing, but actual use of multiple strategies is another. For the questions in this study, the participants used a variety of strategies with the exception of one person who used a search engine consistently, and they were able to articulate reasons for their choices (See Table 2). Eighty percent used all three strategies. An additional 18 percent used two strategies. Usually one strategy dominated, but the range of dominance varied. For those using three strategies, on average, one strategy was used for about half the questions. For those using two strategies, on average, the dominant strategy was used for about two-thirds of the questions.

Perception of Difficulty of Questions

As Table 3 shows, there is a statistically significant difference in the participants= perception of the level of difficulty of the questions based on the type of question. Open questions whose source was unpredictable were considered the most difficult, and closed questions with predictable sources were regarded as the easiest questions. The most difficult and easiest questions represent the opposite types of questions in the typology of questions (See Table 1).

The participants' judgments of the evaluation of the difficulty of the questions support the basis of the typology very clearly. The participants did not know that there were four different types of questions nor did they know the basis of the typology. In the reasons for their judgments, however, they often referred to the factors underlying the typology. Two main reasons given for questions considered easy were that the question was a fact-related question (closed) and that a web-site existed with a strong probability of containing the answer. For example, Finnish participants referred to these reasons for Question 16 (Who is the president of the [Intel or Nokia] company), an easy question:

Table 1. Sample Reference Questions Grouped by Type of Question

	Open/Unpredictable
(1)	What are considered to be the causes of hooliganism or fan violence at World Cup soccer games?
(2)	What international efforts or projects are underway to handle the Year 2000 computer crisis?
(6)	What is the difference between the European approach and the American approach to protecting privacy on the Web?
(13)	What studies are available on the Web about people's knowledge, attitudes, fears, and opinions about virtual reality?
	Open/Predictable
(4)	Amazon Books is often mentioned as a good example of companies whose business is doing well in the Internet. What information is available on the Internet about the company's history and current status, including the kinds of services it offers to customers?
(7)	How is the U.S. Library of Congress resolving copyright issues in connection with its Digital Library project?
(9)	What is the World Health Organization doing to stop river blindness in Africa?
(12)	How do I apply for admission to the medical school at Harvard?
	Closed/Unpredictable
(5)	I need demographic statistics that characterize Internet users -- age, gender, income level, and so on.
(8)	I am looking for a copy of the multinational treaty banning land mines that was signed shortly after Princess Diana's death, the one that the U.S. and Finland refused to sign.
(10)	What does the term "the China Syndrome" refer to?
(14)	Dian Fossey did amazing research studying the habits of gorillas in Africa and became an advocate of maintaining their habitat. She was killed in the course of her research, supposedly by a poacher. Which persons or organizations are continuing her work, if any?
	Closed/Predictable
(3)	What sites are on UNESCO's list of World Heritage sites?
(11)	According to the Bible, how old was Methuselah when he died?
(15)	Who are the current members of NATO, the North Atlantic Treaty Organization?
(16)	Who is the president of the Nokia (Intel in American version) Company in Finland?

Note: Numbers correspond to the order on the questionnaire.

Table 2. Number of Strategies Used and Extent of Dominant Strategy

N of Strategies Used	N of Participants	Concentration in Dominant Strategy	
		Range	Mean
3	43 (79.6%)	38 to 81	48.3 (S.D. 9.6)
2	10 (18.5%)	50 to 88	66.5 (S.D. 11.1)
1	1 (1.9%)	100	N.A.

Notes: Mean N of strategies: 2.78 (SD .46).

Table 3. Level of Difficulty by Type of Question

	Level of Difficulty					
	Difficult		Mid-range		Easy	
Type of Question	N	Percent	N	Percent	N	Percent
Open/Unpredictable	162	75	46	21.3	8	3.7
Open/Predictable	15	6.9	107	49.5	94	43.5
Closed/Unpredictable	84	38.9	104	48.1	28	13
Closed/Predictable	9	4.2	67	31	140	64.8

Note: Chi square = 424.123 with 6 df; Significant at .0001

"This is about the fact. There exists only one correct answer."
"Information is available on the company's homepage. The address of the homepage is easy to be reasoned also by myself."
"Fact information should be found easily, especially when on the web-sites of big organizations usually is a lot of information about their activities."
"The official web site of the Intel Corporation would be the most authoritative source for the information"
"Specific well known organization and public information needed about that organization."

When the participants gave reasons for considering open questions with an unpredictable source to be the most difficult, they referred not only to the broad nature of the questions and unknown sources but also to the overload of information and the possibility of retrieving irrelevant material. Participants described the difficulty of Question 13 (What studies are available on the Web about people's knowledge, attitudes, fears, and opinions about virtual reality?) as follows:

"Broad and difficult question. As a result I should need to check a huge set of web-sites."
"I do not know where to start. There maybe too much information to be found while finding relevant web-documents is difficult."
"Not very common subject expect keywords 'virtual reality', which are very common in www. This means irrelevancy of matches."
AAsking about illegible studies, attitudes, things hard to measure -- no specific organization or group mentioned, no date.@
AHard to search key words on web and have to be careful of sources--might get a lot of hits but not what you want.@
AThe source of this information is too obscure. The question uses broad concepts -- no specific terms to use as a starting point.@

In the web environment, difficulty of questions is clearly related to searchers= perception of the openness of the questions and the unpredictability of the source. In summary, the participants discerned a range of difficulty among the questions in the study and, in their reasoning, indicated that they were influenced by the same factors that had been incorporated into the typology established by the researchers.

Selected Search Strategies

As Table 4 shows, the participants varied their choice of initial web search strategy according to the type of question. When the source was predictable, the searchers in most cases planned to start their search by going straight to the direct address of the certain organization, for example, "How is the U.S. Library of Congress resolving copyright issues in connection with its Digital Library project?" or "Who is the president of the [Intel or Nokia] Company?". If the source was unpredictable, they preferred other strategies.

For open questions, such as "What studies are available on the Web about people's knowledge, attitudes, fears, and opinions about virtual reality?", there were no large differences in the frequency of the selection of directories or search engines, but, if the question was closed and the source unpredictable such as, "What does the term 'the China Syndrome' refer to?", the participants clearly preferred search engines (See Table 4).

Table 4. Search Strategy by Type of Question

	Search Strategy					
	Direct Address		Directory		Search Engine	
Type of Question	N	Percent	N	Percent	N	Percent
Open/Unpredictable	4	1.9	83	38.4	129	59.7
Open/Predictable	145	67.1	27	12.5	44	20.4
Closed/Unpredictable	9	4.2	64	29.6	143	66.2
Closed/Predictable	103	47.7	57	26.4	56	25.9
Total	261	30.2	231	26.7	372	43.1

Note: Chi square = 334.753 with 6 df; Significant at .0001

Direct address, presumably to the Library of Congress web site, was the most often selected strategy to find an answer to the question: AHow is the U.S. Library of Congress resolving copyright issues in connection with its Digital Library project?@ The participants were sure that the answer could be found there. In addition, they considered the homepage to be a reliable source. Participants explained their decisions to go straight to the homepage, for example, saying:

> "I would find the direct address right away (I know a place where I could get it), on its own sites there probably is information about this."
> "I think that if I could get to the web-sites of the Library of Congress, I would find there a link to the web-site of the Digital Library project. And I would find an answer there."
> "[The homepage of U.S Library of Congress is] ... Reliable source."
> "I know the LOC address, and surfing it is simpler/easier than wading through quasi-official sites found by a directory or search engine."
> "Well known particular organization and is about a major project of that organization."

Similarly going straight to the company=s homepage was the most often selected search strategy for the question "Who is the president of the [Intel or Nokia] Company?", a closed question with a predictable source. The participants were sure that the direct address would be the fastest and easiest way to find an answer. They emphasized simple searches and trusted the company's web-site would contain the information. Participants explained their decision to start a search with a direct address for example saying:

> "The company's own web-sites are the best information source."
> "Straight to Nokia's web-sites. Search engine would find too much information about Nokia."
> "Fact about Nokia. The company's own web-sites are the simplest way."
> "Fast and easy."
> AThe official web site of the Intel corporation would be the most authoritative source for this

> information.@
> ASpecific question [about Intel]. Information available from specific source.@
> AIntel page should have the latest information.@

It is not surprising that the participants decided to start their search with a direct address if the source was predictable and if they knew or assumed that a certain web-site existed. Why should a searcher use a map (subject directory) or ask for help (use search engines) if he knows where to obtain what he needs (direct address)? However, the web sites accessed directly are not always easily searchable and/or do not include the desired information.

When the source was unpredictable, the participants did not plan to go to a direct address but preferred other strategies. When the source was unpredictable and the question was closed, participants usually selected (66.2 percent) a search engine initially. They explained their selection by referring both to the clear and specific search terms and to the search engine's suitability for this kind of search. Some reasoned for selecting a search engine for the question, AWhat does the term >the China Syndrome= refer to?@ by saying:

> "Possibility of giving exact search keys."
> "Clear search terms to be used."
> "This is the easiest way to handle single terms."
> "Even though 'The China Syndrome' is a specific term, I am not certain what field it refers to, so I'd want to search a broad range of sources (Ruling out directory and direct address)."
> "I assume any link provided by a search would supply this information (at least briefly or in context). If I don't know the subject in question, I can't use a directory."
> "Greater recall and access to types/contexts using the phrase."

If the source was unpredictable but the question was open, the participants preferred search engines (59.7 percent). Nevertheless, they also selected subject directory more often for this type of question than for other types (38.4 percent). The participants selected a subject directory for their search strategy if they thought that a good subject category existed for this topic in a directory, if they wanted to browse web-sites, or if they found the question to be too complicated for a query formulation. For example, some participants who decided to start their search with a subject directory for the question "What studies are available on the Web about people's knowledge, attitudes, fears, and opinions about virtual reality?" described their reasons as follows:

> "There must be a directory for a virtual reality."
> "Unclear search, with a subject directory I could get started."
> "I use a subject category because I cannot formulate a query which would be specific enough to be used by search engines. Category is Computers & Internet/WNW."
> AI could look under whatever heading most closely resembles Computer Science Virtual Reality Studies and skip all the games and vendors and other garbage a search engine would give me.@
> AA directory would give ideas leading to what category might best answer this question. It would be hard to search for and there is no way to find a direct address.@

Reasons given for choosing a search engine for the same search usually referred to the words mentioned in the question and to the possibility of combining words in a search statement, but also to their trust in search engines. They commented:

"Clear search terms."
"In the question there are many good search terms."
"A search engine would find more information about this topic."
AWide-ranging question with numerous, unpredictable sources. Do not want to limit search.@
ALarge topic. Would require different sites, widely scattered.@

On the basis of the data participants use different reasons for choosing between a subject directory and a search engine. The reasons for selecting a search engine seem to be related to the terms mentioned in a question. If the question is closed or otherwise includes exact terms, it supports the selection of a search engine. In the cases where searchers find a question broad, they seem to prefer subject directory as the initial strategy.

Discussion

The general perception about web users that has emerged from previous studies is that they are not very sophisticated searchers: preferring known sites, browsing, using only simple searches if they use a search engine, and sometimes being relatively non-discriminating in recognizing relevant information. The complexity of the individual search may induce these coping skills. On the other hand, the findings of this study begin to suggest another perception when the options are relatively few. The participants in this study not only indicated a fairly high degree of familiarity with their initial search options and used multiple search strategies but also said they would be influenced in their choice of an initial search strategy by question-related variables. They seem to be matching the capabilities of the search strategies with the requirements of the question. This study, of course, did not ask the participants to actually search the questions, and it is possible that in the real world actually searching may not match questionnaire responses.

Another important finding is that the participants agreed to a large extent on the level of difficulty they attached to each question. Analyzing the reasoning behind their decisions about difficulty, as will be done in a subsequent paper, should elucidate in greater detail the factors they considered in making these determinations to see if they agreed on reasoning as well. The examples included in the paper seem to indicate that they are at least considering the notion of constraints on the answer created by open and closed questions and their ability to predict the source of the answer.

These findings have implications for teaching people about the web. First, web users should be taught the strengths and weaknesses of each approach available to them at the first stage of a search. Many seem to be able to extrapolate these from their experiences to develop this framework, but incorporating it explicitly into the canon of web instruction would insure that all users have a similar and detailed perspective on these approaches. Not all users may be able to draw inferences from their experiences nor may their experiences be broad enough so that they understand the full range of strengths and weaknesses. As we discussed in the introduction, there are clear differences between direct address, subject directories, and search engines as initial search strategies. In addition, there are also clear differences between them when knowledge pre-requisites are considered. If a user does not maintain a list of known sites, going directly to a certain web page (direct address), calls for her to be able to identify likely sources of information, to know how the address is structured, to which category the organization belongs, e.g. eau or com, and how abbreviations are written. The knowledge pre-requisites for the others are different and, in some cases, perhaps even more demanding.

In addition, question analysis should be addressed in detail. In many cases, understanding the full implications of the question and its anticipated appropriate answer would provide insights into the use of each of these options. In directory use, for example, it would allow the user to categorize his subject and move more fluidly among the classified subjects in the directory; in search engines, it should facilitate the selection of search terms.

Another important implication of these findings is that researchers about web searching should consider question-related variables (independent of the particulars of the question) as an influential factor in search decisions and begin to consider and hopefully to elucidate the influence that they have on search decisions. Failing to consider explicitly question-related variables is folly since their hidden influence may confound other findings.

References

Fidel R. et al. (1999). A visit to information mall: Web searching behavior of high school students. *JASIS - Journal of the American Society for Information Science* 50 (1), 24-37.

Hsieh-Yee, I. (1998). Search tactics of Web users in searching for texts, graphics, known items and subjects: A search simulation study. *Reference Librarian* 60, 61B85.

Siegel, S. & Castellan, N.J. (1988). *Nonparametric statistics for the behavioral sciences*. (2nd ed.). New York: McGraw-Hill.

Spink, A., Bateman, J. & Jansen, B.J. (1998). Searching heterogeneous collections on the Web: Behaviour of Excite users. *Information Research : An electronic journal* [online], 4 (2). Available: URL:http://www.shef.ac.uk/~is/publications/infres/paper53.htm.

Tillotson, J., Cherry, J. and Clinton, M. (1995). Internet use through the University of Toronto Library: Demographics, destinations, and users' reactions. *Information Technology and Libraries* 1, September, 190B198.

Wang, P., Tenopir, C. (1998). *An exploratory study of users' interaction with World Wide Web resources: Information skills, cognitive styles, affective states, and searching behavior.* Contributed paper to Annual National Online Meeting 1998, New York, NY.

Wang, P. et al. (1998). *An exploratory study of user searching of the World Wide Web: A holistic approach.* Contributed paper to ASIS Annual Meeting 1998, Pittsburgh, PA.