Observations of Searchers: OHSU TREC 2001 Interactive Track

William Hersh
Lynetta Sacherek
Daniel Olson

Division of Medical Informatics & Outcomes Research
Oregon Health & Science University
Portland, Oregon, USA

{hersh, sacherek, olsondan}@ohsu.edu

Abstract

The goal of the TREC 2001 Interactive Track was to carry out observational experiments of Web-based searching to develop hypotheses for experiments in subsequent years. Each participating group was asked to undertake exploratory experiments based on a general protocol. For the OHSU Interactive Track experiments this year, we chose to perform a pure observational study of watching searchers carry out tasks on the Web. We found users were able to complete almost all the tasks within the time limits of the protocol. Future experimental studies aiming to discern differences among systems may need to provide more challenging tasks to detect such differences.

Background

At the SIGIR 2000 workshop, Interactive Retrieval at TREC and Beyond, the Interactive Track decided on a number of new directions for TREC 2001 and beyond [1]. One of these decisions was to move the track to a two-year cycle, which would provide time to refine the track protocol and collect adequate amounts of data. In addition, it was agreed upon that the TREC 2001 track activities would consist of observational studies aiming to view searchers in realistic searching situations and to generate hypotheses that could assessed when the track returned to experimental studies in the following year.

For TREC 2001, participants in the Interactive Track carried out observational studies that aimed to maximize the realism of the searching by allowing the use of data and search systems/tools publicly accessible via the Internet. Within the framework of a broad protocol, groups were encouraged to allow their searchers to choose tasks and systems/tools for accomplishing those tasks. Groups were also asked, however, to maximize the likelihood they would find in their observations a hypothesis they could test for TREC 2002.

The OHSU group chose to undertake a purely observational study for TREC 2001, watching searchers carry out the assigned tasks with Web tools they desired to use. Users were allowed to choose whatever tools they desired for each questions, with their actions logged via the browser history files. They were also administered a short questionnaire after each question about their knowledge of the topic, ease of searching, and perceived success of searching.

Methods

The OHSU Interactive Track group performed an observational study that adhered to the guidelines of the track protocol. Each subject was asked to perform a search for each of four domains:

1. Finding consumer medical information on a given subject
2. Buying a given item
3. Planning travel to a given place
4. Collecting material for a project on a given subject

Per the track protocol, two of the questions would be fully-specified (i.e., user given the complete task) and the other two would be partially-specified (i.e., the user would decide part of the task). As the protocol had four domains, two search types (fully-specified and partially-specified), and two questions for each domain and type, a total of 16 questions were available (see Table 1). For data analysis, a three letter/number identifier was developed to identify each question based on the domain, type, and question number. The questions and their order were permuted for each searcher to insure each searcher would have one each from the four domains that varied in terms of type, question number, and the order administered.

Searchers were asked to search on each question as they would normally search on the Web, using the search engines, catalogs, etc. that they would typically use. They were instructed to provide a pertinent answer to each question along with the URL(s) to justify it. After they finished searching, they were asked to answer five questions on a Likert scale of 1 (not at all) to 5 (extremely). These statements were:

1. Are you familiar with this topic?
2. Was it easy to get started on this search?
3. Was it easy to do the search on this topic?
4. Are you satisfied with your search results?
5. Did you have enough time to do an effective search?

The searchers were given up to 20 minutes to complete each question but were asked to stop if they completed it sooner, with the elapsed time recorded. Each Web page they viewed was tracked by the Netscape History file, which we cleared before the searcher arrived and saved when the session was over.

For each question, we obtained a number of variables. We first determined whether the task was completed successfully. We did not check whether the user answered the question "correctly," but rather determined whether he or she listed an appropriate answer (or number of answers) pertinent to the task. We also measured the time taken for the question, the number of pages viewed, and the number of pages listed as justifying the answer.

Similar to previous track years, we recruited experienced searchers who were librarians or information professionals in the Pacific Northwest or medical informatics graduate students at OHSU. The experiments took place in computer training rooms at OHSU where searchers could choose to use either a Windows or Macintosh computer connected to the Internet running Netscape Navigator. Searchers were given a ten-minute orientation describing the nature of the experiment and the plan for the two-hour session. They also performed a practice search which asked them, "Find universities that offer a graduate degree in computer science in Oregon."

Because our experiment was exploratory in nature, we had no a priori hypotheses and thus performed no statistical analyses.

Results

A total of 24 searchers were recruited. All were highly experienced Web searchers, either as information professionals or graduate students. All subjects completed with experiments without difficulty. Four questions were discarded due to incomplete data collection, so a total of 92 questions were analyzed. A wide variety of topics were chosen for the partially-specified questions, as shown in Table 2.

Table 3 shows the overall results at the per-question level and two levels of aggregation. As seen in the first section of Table 3, virtually all subjects completed the tasks correctly. The average time taken was well under the 20 minutes allowed. The lower sections of Table 3 show aggregation of the results by domain-search type and further by domain. The differences across domains are small, but the buying task did take the most time and required the most page views.

Table 4 shows the results of the post-searching questions. While searchers were equivocal about their prior knowledge of the topic, they were consistent in their high belief that the search was easy to do and provided satisfactory results within an adequate amount of time.

We also looked at the use of search engines. A total of 68 questions (74%) employed the use of one search engine. Sixteen questions (17%) used no search engines, while seven (7%) employed two different ones and one (1%) used four. Table 5 shows the search engines used. Google was the most commonly used search engine, with Yahoo being the only other one used more than five times.

Further analysis planned after the TREC meeting for the final proceedings paper will focus on analyzing the search engines and catalogs used along with the queries posed to them.

Conclusions

Our study demonstrates that experienced Web users are able to perform relatively challenging tasks successfully using the search engines and catalogs that are available. This result is not new, but does demonstrate that such information tasks can be performed by a wide cross-segment of experienced and educated users.

This study does not provide any major new insights into user searching, but it clearly does indicate that experiments aiming to discern differences among systems will need to employ tasks that are more difficult. Otherwise users will be able to complete tasks no matter what systems they use, and any differences among systems will not be detected. This study also shows that users employ a variety of high-quality, widely-available tools in their searching. Experiments that focus on single systems or user interfaces may not provide the diversity of approaches that would accommodate all searchers.

References

1. Hersh W, Over P, SIGIR workshop on interactive retrieval at TREC and beyond, SIGIR Forum, vol. 34, no. 1.

Table 1 - Searching questions with domain (medical, buying, travel, or project), search type (fully-specified or partially-specified), number, and text. Blank line indicated area for user choice in partially-specified questions.

Domain	Search Type	Question Number	Question Text
M	F	1	Tell me three categories of people who should or should not get a flu shot and why.
M	F	2	Find a website likely to contain reliable information on the effect of second-hand smoke.
M	P	1	List two of the generally recommended treatments for _____.
M	P	2	Identify two pros or cons of taking large doses of _____.
B	F	1	Get two price quotes for a new digital camera (3 or more megapixels and 2x or more zoom).
B	F	2	Find two websites that allow people to buy soy milk online.
B	P	1	Name three features to consider in buying a(n) _____.
B	P	2	Find two websites that will let me buy a(n) _____ online.
T	F	1	I want to visit Antarctica. Find a website with information on organized tours/trips there.
T	F	2	Identify three interesting things to do during a weekend in Kyoto, Japan.
T	P	1	Identify three interesting places to visit in _____.
T	P	2	I'd like to go on a sailing vacation in _____, but I don't know how to sail. Tell me where can I get some information about organized sailing cruises in that area.
P	F	1	Find three articles that a high school student could use in writing a report on the Titanic.
P	F	2	Tell me the name of a website where I can find material on global warming.
P	P	1	Find three different information sources that may be useful to a high school student in writing a biography of _____.
P	P	2	Locate a site with lots of information for a high school report on the history of _____.

Table 2 - Topic designations for partially-specified question.

Searcher ID	Question ID	Partial Topic Text
4	BP1	TV
5	BP1	SUV
12	BP1	DVD Player
13	BP1	bicycle
20	BP1	house
21	BP1	hybrid car
3	BP2	Laptop computer
7	BP2	book
11	BP2	airplane
15	BP2	bassinet
19	BP2	set of dishes
23	BP2	(4 door) car-compact
3	MP1	Heart Disease/Atrial fibrillation
8	MP1	Achalasia
11	MP1	eczema
16	MP1	breast cancer
19	MP1	sunburn
24	MP1	scabies
2	MP2	Steroids
10	MP2	pseudoephedrine HCl
14	MP2	aspirin
18	MP2	Vitamin C
22	MP2	vitamin C
2	PP1	Shakespeare
7	PP1	Abraham Lincoln
10	PP1	Virginia Woolf
15	PP1	Amelia Earhart
18	PP1	Ayn Rand
23	PP1	Mark Twain
1	PP2	Turkey
4	PP2	cat worship
5	PP2	the Great wall of China
9	PP2	Mesopotamia
13	PP2	computers
17	PP2	Eleanor Roosevelt
21	PP2	WWII
1	TP1	Santa Fe, NM
9	TP1	Greece
14	TP1	Burma
17	TP1	Berlin
22	TP1	Oregon
4	TP2	Tahiti
8	TP2	wales
12	TP2	Hawaii
16	TP2	The Galapagos
20	TP2	the Bahamas
24	TP2	San Juan Islands

Table 3 - Overall results by individual question and averaged by full-partial status and domain.

Question ID	Number of Searchers	Number Correct	Percent Correct	Time (minutes)	Pages Viewed	Pages Listed
By Individual Question						
BF1	5	4	80%	11.00	15.20	2.00
BF2	6	6	100%	9.50	29.83	2.17
BP1	6	6	100%	7.00	15.17	3.17
BP2	6	6	100%	8.67	15.50	2.67
MF1	6	6	100%	8.50	9.50	2.00
MF2	6	6	100%	5.83	10.00	2.33
MP1	6	5	83%	6.00	11.17	1.50
MP2	5	4	80%	11.80	12.60	2.00
PF1	5	5	100%	5.80	12.40	2.80
PF2	5	5	100%	3.80	5.80	2.40
PP1	6	6	100%	8.17	12.17	3.17
PP2	7	7	100%	7.14	13.29	2.29
TF1	6	6	100%	6.00	13.83	1.17
TF2	6	5	83%	7.50	15.00	1.67
TP1	5	5	100%	6.80	14.20	2.80
TP2	6	6	100%	7.83	12.00	3.00
By Full-Partial Question						
BF	5.5	5	91%	10.25	22.52	2.08
BP	6	6	100%	7.83	15.33	2.92
MF	6	6	100%	7.17	9.75	2.17
MP	5.5	4.5	82%	8.90	11.88	1.75
PF	5	5	100%	4.80	9.10	2.60
PP	6.5	6.5	100%	7.65	12.73	2.73
TF	6	5.5	92%	6.75	14.42	1.42
TP	5.5	5.5	100%	7.32	13.10	2.90
By Domain						
B	5.75	5.5	96%	9.04	18.93	2.50
M	5.75	5.25	91%	8.03	10.82	1.96
P	5.75	5.75	100%	6.23	10.91	2.66
T	5.75	5.5	96%	7.03	13.76	2.16

Table 4 - Questionnaire results by individual question and averaged by full-partial status and domain.

Question ID	Familiar with topic	Easy to get started	Easy to do search	Satisfied with results	Enough time to search
By Individual Question					
BF1	2.80	5.00	4.80	3.80	4.80
BF2	2.17	4.33	3.67	4.33	4.83
BP1	2.33	4.50	4.50	4.67	4.83
BP2	3.17	4.67	4.50	4.33	4.83
MF1	2.67	4.83	4.33	4.83	5.00
MF2	2.67	4.33	4.17	4.67	4.83
MP1	3.33	3.83	3.83	4.17	4.00
MP2	3.20	4.20	3.80	3.40	4.00
PF1	2.80	4.60	4.60	4.40	4.80
PF2	2.60	4.40	4.80	4.20	4.80
PP1	3.17	4.83	5.00	4.67	4.67
PP2	3.00	4.71	4.43	4.43	4.71
TF1	1.17	3.83	3.83	3.83	4.67
TF2	1.67	4.67	4.33	4.33	4.67
TP1	3.00	4.60	4.80	4.80	5.00
TP2	1.67	4.33	4.33	4.33	4.67
By Full-Partial Question					
BF	2.48	4.67	4.23	4.07	4.82
BP	2.75	4.58	4.50	4.50	4.83
MF	2.67	4.58	4.25	4.75	4.92
MP	3.27	4.02	3.82	3.78	4.00
PF	2.70	4.50	4.70	4.30	4.80
PP	3.08	4.77	4.71	4.55	4.69
TF	1.42	4.25	4.08	4.08	4.67
TP	2.33	4.47	4.57	4.57	4.83
By Domain					
B	2.62	4.63	4.37	4.28	4.83
M	2.97	4.30	4.03	4.27	4.46
P	2.89	4.64	4.71	4.42	4.75
T	1.88	4.36	4.33	4.33	4.75

Table 5 - Frequency of use of search engines.

Search Engine	Times used
Google	45
Yahoo	20
Ask Jeeves	5
Alta Vista	4
Ask	4
Metacrawler	4
Netscape	3
Excite	2
Lycos	2
AOL	1
LookSmart	1