Task behaviors during web search: 
The difficulty of assigning labels

Daniel M. Russell, Diane Tang, Melanie Kellar, Robin Jeffries 
Google 
1600 Amphitheatre Parkway 
Mountain View, CA 94043 
{drussell, diane, kellar, jeffries @google.com}

Abstract
By examining searcher behavior on a large search engine, we have identified seven basic kinds of task behaviors that can be observed in web search session logs. In the studies reported, we first manually labeled 700 complete web sessions, and then subsequently had 23 searchers self-label 252 days of their own sessions to give an accurate picture of what kinds of tasks people are doing when they search. From these two studies, we have found that the most accurate labeling of search task session data is done by the searchers themselves, and that it is very difficult for an external observer or automatic classifier to infer where the task boundaries are or what the actual user task goal is.

1. Introduction

How do searchers behave while doing web search? Do they unerringly create a query, refine it, and discover the answer to their question? The more we understand about user intent and behaviors, the better we can tune search engine algorithms and the user experience of search. However, understanding user intent is difficult; without knowing the task a priori, examining a web search session requires a leap of understanding and empathy with the searcher, since a common reaction is one of amazement—“what is this person trying to do?”

Related work: There has been a great deal of previous research that examines and identifies categories of information seeking and searching behavior. The methodology used to develop these categorizations usually consists of log analysis combined with a secondary method for discovering the user intent. This secondary method ranges from qualitative interviews, to self-labeled data, to third-party ratings and classifications. [2, 4, 6, 7, 8]

These categorizations were developed by examining the entire sequence of user actions. In contrast, research examining search goals and queries is often conducted through the analysis of single queries, often without the context of a session or user-specified intent. Table 1 provides an overall view of the distribution of search query classification. Broder [1] developed a taxonomy of web search goals using both randomly selected single search queries and an analysis of survey data collected from AltaVista users.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Survey</td>
<td>Log</td>
<td>Set 1</td>
</tr>
<tr>
<td>Navigational</td>
<td>25%</td>
<td>20%</td>
<td>15%</td>
</tr>
<tr>
<td>Informational</td>
<td>~39%</td>
<td>48%</td>
<td>63%</td>
</tr>
<tr>
<td>Resource/Transaction</td>
<td>~36%</td>
<td>30%</td>
<td>22%</td>
</tr>
</tbody>
</table>

Table 1: Reported distributions for search query categorizations

The taxonomy includes three categories of search goals: navigational, informational, and transactional (where the user’s goal is to perform a web-based activity). Building on Broder’s work, Rose and Levinson [9] developed a search goal hierarchy, consisting of three top-level search goals: navigational, informational and resource.

Jansen, Booth, and Spink [5] examined the session characteristics of navigational, informational, and transactional search queries. These characteristics were used to develop a classification algorithm that was validated with 400 manually classified single queries and achieved an accuracy of 74%.
2. Two Studies of Search Task Behavior

2.1 Study #1
Identifying New Task Categories

We initially tried having humans assign search-task-type labels to search sessions using the Rose and Levinson taxonomy. In doing so, we found that there is often great ambiguity; while we found good inter-rater agreement on navigation search type sessions, we were not able to get sufficient inter-rater agreement on more ambiguous search tasks. As a result, the goal of our first study was to attempt to reproduce the earlier search task categorization work without assuming their task categories.

Method: In mid-2006 we began an effort to re-derive a canonical set of search-task descriptions. In a long series of analyses, a group of seven web research professionals reviewed 700 anonymized sessions. Sessions were randomly sampled from a world-wide set of English-language sessions during March-April 2006, and varied in length from 2 to 100 events (1 to 5 separate search tasks) and spanning no more than 24 hours. Thus, each session comprised all the activity for a given cookie during a single day. Each session was carefully anonymized according to policy, and contained the queries, query refinements, next pages, the complete list of results shown, and the result clicks.

Each session was closely read by each researcher, then discussed to determine searcher behavior and intent. The panel attempted to find labels that would accurately characterize each session. After the first 100 sessions were open-coded in this way, definitions of each label were written by the panel, then applied to the next set of 100 sessions. Each iteration lasted approximately one day, during which new labels were proposed as needed, label definitions rewritten, and ill-fitting labels discarded or reworked.

At the end of the process, the panel went through 7 iterations, emerging with a new search-task categorization made up of session behavior labels and definitions. As in other analyses of this kind [1, 2, 5, 9], this panel identified 7 basic kinds of search-task behaviors, although our basic task definitions vary from previous literature. (See Table 2.)

<table>
<thead>
<tr>
<th>Task Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigate</td>
<td>query(ies) leading to a site at which the main task can be performed. Navigation queries are often the names of destination sites, or terms that are heavily suggestive of the target site. Example: find Apple’s Quicktime web page.</td>
</tr>
<tr>
<td>Find-Simple</td>
<td>searching for an evident piece of information that does not require multiple sources of information. Example: find the local YMCA phone number.</td>
</tr>
<tr>
<td>Find-Complex</td>
<td>searching for information that requires searches on very closely related topics to integrate information across resources; often marked by a sequence of searches to find information that is coordinated. Example: search for a way to knit winter socks that uses a search for sock patterns and knitting instructions</td>
</tr>
<tr>
<td>Locate.Acquire</td>
<td>the searcher is looking to download something, purchase an item, obtain a good or service. Example: find tickets to the game.</td>
</tr>
<tr>
<td>Explore.Learn</td>
<td>searches that are intended to discover something or learn about a topic area. Example: finding out how to raft on the Grand Canyon, information about the tourist season, local flora and fauna.</td>
</tr>
<tr>
<td>Play</td>
<td>activity where the searches are intended to find games, fun content or items for amusement. Example: finding sites to play online computer games; celebrity searches, etc.</td>
</tr>
<tr>
<td>Meta</td>
<td>search tasks that are to test some capability. Example: testing a web site’s capabilities.</td>
</tr>
</tbody>
</table>

Table 2: Web search task definitions
2.2 Study #2

Now armed with a new search-task taxonomy, we began asking for users to create their own labeled web search histories.

Method: To create a new search session sample, 23 software professionals (a mix of software engineers and UX practitioners) at Google were recruited for this study. (They were not compensated for their participation.) Each subject was given instruction on how to segment and label their personal web history and instructed to take their time and strive for accuracy rather than quantity. Each subject labeled 1 to 2 weeks of their immediate past web search history, dividing the history into labeled tasks and indicating both the number of tasks performed and the number of searches each task required. For anonymity, we did not ask for the log itself, but just the day, the sequence, the task label, and the number of searches done for each task.

We relied on the participants to self-identify the task boundaries and task types based on our taxonomy description. It quickly became clear that in many cases, the searcher had access to internal state and information that wasn’t clearly available to an external observer, such as where the task boundaries were between very similar-appearing searches.

Analysis: A few observations became immediately apparent. First, even though this user population is highly biased towards sophisticated users, the total number of tasks and searches / user / day was less than we had expected.

| N searchers | 23 |
| Number of days of tasks | 252 |
| Number of labeled tasks | 957 |
| Number of searches | 2109 |
| Searches / user / day | 8.05 |
| Tasks / user / day | 3.80 |

Table 4: Distribution of task types from sample.

It’s clear that, in aggregate, searching is relatively efficient in this population: people need only a small number of searches to accomplish a task (~2). In fact, tasks with 1, 2, or 3 searches were 84% of all tasks, with 55% of all tasks involve only 1 search, 18% have 2 searches, and 10% have 3.

During the process of creating the taxonomy, we were tempted to further divide up the FC and FS categories, especially given that they are the most common task type as seen in Table 4. However, during the data collection process, we fielded enough queries about the exact definitions for FS vs. FC that we became unconvinced that even searches themselves, much less third party raters, would be able to more accurately categorize their behavior.

Figure 1: Find-Simple (FS) and Find-Complex (FC) tasks take different numbers of searches to reach their goals.

However, when we compare the task length of FC and FS tasks, it is clear that FS tasks really are simple: 66% of all FS tasks take only 1 query while 62% of all FC tasks take 2 or more. On average, FS tasks require 1.6 searches / task, while FC tasks need 2.6 searches. As Figure 1 shows, complex tasks require more searches (and, we suspect, more time, although this data does not include that fine resolution time-to-result data).

3. Discussion

Labeling tasks is hard: Our first study looked for a set of task labels that we could accurately use to label searcher behavior. While earlier work suggested there is a fine set of distinctions to be made, we quickly learned that the application of a fine-grained set of labels was problematic in practice. In Study #1 we found that even trained raters had a
difficult time following analysis instructions and creating a consistent labeling of web search session behaviors. So we turned to having the searchers label their own behaviors and gathered a reasonable sized dataset from a highly skilled and frequent searcher population.

Population bias: Even with such a bias, the actual behaviors seem somewhat surprising, with fewer tasks per day than expected (~3 tasks / user / day). Those tasks tend to be Find-Complex or Find-Simple in nature, and tend to be fairly efficient, taking only 1, 2, or 3 searches to successfully complete.

The relative ratio of task types was also somewhat surprising, with FC leading the overall tasks, and making up just over 1/3rd of all tasks. We suspect that the FC is high because of user population in this study. FC tasks pull together information from a number of sources, often (but not always) from multiple web pages to resolve a question that doesn’t have a single, findable answer. This is a common task type for this population. Similarly the number of Play tasks (P) seems low by contrast to other reports. We hypothesize that different populations will have rather different relative task profiles. Not all populations will have high FC and low P. Because of these differences, we would encourage other researchers to include in similarly detailed population descriptions in their results.

4. Comparison with literature

Although comparisons are somewhat difficult between different studies in the literature, we noted some differences between what the Rose and Levinson study reported and what we found. In their sample, ~60% of searches are informational, ~13% navigational, and 24% resource access. Nearly all of the studies agree that Navigation comprises ~15% of all search tasks, including ours. The one exception is from Broder, where the survey shows 25% navigational [1]. We note that this number was derived by a pop-up survey of AltaVista users in the middle of their search task. It’s unclear if the questions were rotated to avoid first-order question effects, but given that the logs analysis of Navigation queries in the same study showed 20%, it would seem that both numbers are high with-respect-to current practices.

We label the majority tasks as either Find-Complex (FC) or Find-Simple (FS), while Rose [9] divides the tasks into subcategories of Directed, Closed/Open, Advice / Locate / List. While these seem clear enough when seen example by example, we found the distinctions too dependent on the details of the user’s mental state at the time of the task to be determined externally. It was illuminating to us that our human task labelers kept asking questions to refine the definition of EL vs. FC and FS vs. FC. To compensate for this problem, we shifted our taxonomy to be based on a simpler criterion to keep rating accuracy high.

If we group FC and FS together and compare directly to the Informational category of Rose, then we see rough agreement: 57.3% (our study) vs. ~61% (Rose). By contrast, Jansen [5] rates Informational tasks at ~81%, but we believe their classifier is a bit too broad (queries are classified as Informational if they contain one or more of a set of informational cue terms, which is probably too open for this kind of classifier).

5. Summary

Inferring user intent from their behavior is more difficult that it would first appear. Yet knowing their intents and goals is essential in assessing whether or not the user interaction for search is working well or not.

And while tasks are predominantly short (<3 searches according to our population, and this observation is consistent with the population at large), labeling the task requires knowing significantly more about the goal and mental state of the searcher than can be determined just by examining the query. Even apparently simple queries can hide complexities.

More generally, inferring user behavior from a small number of data points is very hard. This is a lesson that applies generally to web-based interactions. Task boundaries are often difficult to determine and the underlying goals are even harder.

This suggests that the approach to getting high quality labeled behavior data is to have the searchers do self-rating. In the face of our inability to get reliable inter-rater results and the complexity of web search trace as we see it in the logs, this will likely be the case for quite some time to come.
6. References


