Logging the Search Self-Efficacy of Amazon Mechanical Turkers

Henry Feild*
Dept. of Computer Science
University of Massachusetts
Amherst, MA 01003
hfeild@cs.umass.edu

Rosie Jones
Yahoo! Labs
4 Cambridge Center
Cambridge, MA 02142
rosie.jones@acm.org

Robert C. Miller
MIT CSAIL
32 Vassar St
Cambridge, MA 02139
rcm@mit.edu

Rajeev Nayak
MIT CSAIL
32 Vassar St
Cambridge, MA 02139
jeev@mit.edu

Elizabeth F. Churchill
Yahoo! Research
4301 Great America Parkway
Santa Clara, CA
elizabeth.churchill@yahoo-inc.com

Emre Velipasaoglu
Yahoo! Labs
701 First Ave
Sunnyvale, CA 95054
emrev@yahoo-inc.com

ABSTRACT

Conducting focused but large-scale studies and experiments of user search behavior is highly desirable. Crowd-sourcing services such as the Amazon Mechanical Turk allow such studies to be conducted quickly and cheaply. They also have the potential to mitigate the problems associated with traditional experimental methods, in particular the relatively small and homogenous participant samples used in typical experiments. Our current research project addresses the relationship between searcher self-efficacy assessments and their strategies for conducting complex searches. In this work-in-progress paper, we describe our initial tests of using Amazon Mechanical Turk to conduct experiments in this area. We describe a platform for logging the actions taken by Turkers, and a questionnaire we conducted to assess search self-efficacy of average Turkers. Our results indicate Turkers have a similar range of search self-efficacy scores to undergraduate students, as measured by Kelly [8]. We were able to reach a large number of searchers in a short time and demonstrated we can effectively log interactions for rigorous log-based evaluation studies. Changing the amount of remuneration Turkers received had a significant effect on the time spent filling out the questionnaire, but not on the self-efficacy assessments. Finally, we describe the design of an experiment to use Turkers to evaluate search assistance tools.

*Work completed while at Yahoo!

SIGIR'10 Workshop on Crowdsourcing July 19–23, 2010, Geneva, Switzerland.

General Terms

Keywords

search, self-efficacy, crowd-sourcing, Amazon Mechanical Turk, user study, iterative method design

1. INTRODUCTION

This work is motivated by our current research project to study the effectiveness of search assistance. Following Compeau and Higgin's [5] research suggesting that a determiner of computer use is people's self-efficacy assessments around computer literacy or competence, our hypothesis is that search self-efficacy may affect users' willingness to interact with search assistance tools such as relevance feedback and query suggestions. Self-efficacy is defined by Bandura to be "concerned with people's beliefs in their capabilities to produce given attainments" [2]; self-efficacy measures offer an assessment of a person's confidence in their ability to perform task(s). We would like to study users with a range of search self-efficacy levels, and log their interactions with a search engine, including a variety of search assistance tools.

We have several desiderata when we attempt to evaluate the quality of search engines for web search users. Firstly we would like to evaluate over a representative sample of search users. An effective way of doing this is with live tests on a search engine such as described by Anick [1]. However, live tests have two draw-backs: they are risky in that a bad test could alienate users. In addition, the meaning of user click and interaction behavior is still an area of active research, and its relationship to goal success and user satisfaction is still only approximately understood [7, 3].

A second desirable property is to understand the range of users well. Finding study populations in universities allows us to study the users in detail, including surveying their demographics, and other properties, but these users may not be representative of general web searchers. In particular for our study we would like to sample web searchers with a range of search abilities, orientations to Internet use, and search self-efficacy levels. Getting participants who are

representative along all these dimensions is unlikely in an easily accessible, and relatively homogenous population like students at a university or workers in an Internet company. For example, considering our current area of interest, self-efficacy, Kelly [8] measured the search self-efficacy of undergraduate students, and found that they had generally high search self-efficacy. To evaluate search on such a population may overestimate the ease with which people find things, by under-representing low search self-efficacy users.

Running tests on crowd-sourcing services such as the Amazon Mechanical Turk (AMT) may mitigate the problems with both university-based and live-search based evaluations. However, there are two challenges in evaluation using workers on AMT ("Turkers"). The first is logging the searches and clicks they perform during the task. Turkers prefer not to download software or toolbars that could be used to track their interactions. The second is understanding how representative Turkers are of general web searchers.

In Section 2 we describe search self-efficacy in more detail, and give the details of the search efficacy scale we use in this work. In Section 3 we describe our preliminary results from the questionnaire on search self-efficacy of Turkers and compare it to the results obtained by Kelly [8] on undergraduate students. In Section 4 we give details of the platform which we will use to log search interactions in our full study. In Section 5 we describe some open design issues for our full study, arising from this preliminary study of Turkers. Finally in Section 6 we describe the full study we are preparing, which will measure the effectiveness of search assistance tools for searchers with different levels of search self-efficacy, and different levels of frustration.

2. SEARCH SELF-EFFICACY

Kelly's search self-efficacy scale, presented below, covers a range of activities involved in searching, from general query formulation to query refinement to results filtering and management. Users are asked to rate their self-confidence on a number of tasks using a numerical scale from 1 to 10, where 1 is *totally unconfident* and 10 is *totally confident*. Questions on the scale are as follows:

I can...

1. Identify the major requirements of the search from the initial statement of the topic.
2. Correctly develop search queries to reflect my requirements.
3. Use special syntax in advanced searching (e.g., AND, OR, NOT).
4. Evaluate the resulting list to monitor the success of my approach.
5. Develop a search query which will retrieve a large number of appropriate articles.
6. Find an adequate number of articles.
7. Find articles similar in quality to those obtained by a professional searcher.
8. Devise a query which will result in a very small percentage of irrelevant items on my list.
9. Efficiently structure my time to complete the task.
10. Develop a focused search query that will retrieve a small number of appropriate articles.
11. Distinguish between relevant and irrelevant articles.
12. Complete the search competently and effectively.
13. Complete the individual steps of the search with little difficulty.
14. Structure my time effectively so that I will finish the search in the allocated time.

In presenting our results, we use these numbers to reflect efficacy assessments across our participant population.

	$0.50 HIT	$0.05 HIT
Min	47.00	28.00
Median	117.50	92.50
Mean	134.89	99.06
Max	503.00	123.75
Stddev	74.92	50.02

Table 1: Statistics about the time (in seconds) each user spent on the surveys.

3. RESULTS

We ran a questionnaire on AMT two different times, each with 100 Turkers. The questionnaire asked users their age and gender in addition to the fourteen search self-efficacy questions presented in section 2. For our first presentation of the questionnaire we paid workers $0.50 to fill out the questionnaire. The second presentation of the questionnaire paid only $0.05. While the time of month varied, the day of the week and time of day when the AMT human-intelligence task (HIT) was released was the same. We compare the differences between these two presentations below.

3.1 Demographics

The populations showed very similar gender splits and a somewhat similar spread in ages. The workers that completed the first HIT consisted of 57 males and 43 females. Their ages ranged from 18 to 81, with a mean of 32 years. For the second HIT, there were 55 males and 45 females ranging in age from 18 to 62 years old, with a mean of 30.

3.2 Time to Completion

Each HIT was released at 8:30pm American Eastern Daylight Savings Time (EDT) on two different Mondays during June 2010. The $0.50 HIT was released first. Within 106 minutes, all 100 assignments were accepted by workers. The second HIT was issued a few weeks later and it took 540 minutes for all 100 assignments to be accepted—*five times as long*. Table 1 shows the statistics for per-worker survey completion in seconds for each of the survey versions. The means are statistically different according to a Welch's two-sided t-test ($p < 0.0002$). We see that workers spent significantly more time on the questionnaire in the first presentation, when workers were paid $0.50 rather than $0.05.

3.3 Self-efficacy Responses

Users were asked to rate their confidence in being able to perform each of the fourteen search self-efficacy questions using the scale described in Section 2. Figure 1 shows the range of responses for each question for the two presentations of the questionnaire. We can see that both plots are skewed towards the higher end, suggesting a ceiling effect.

The mean over average scores per user was 7.63 (sd=1.38; min=3.74; max=10.00) for the $0.50 version of the questionnaire and 7.26 (sd=1.35; min=3.86; max=10.00) for the $0.05 version. The average scores for our two HITs did not differ significantly according to a two-sided T-test ($p = 0.054$). The scores seem consistent with the mean found

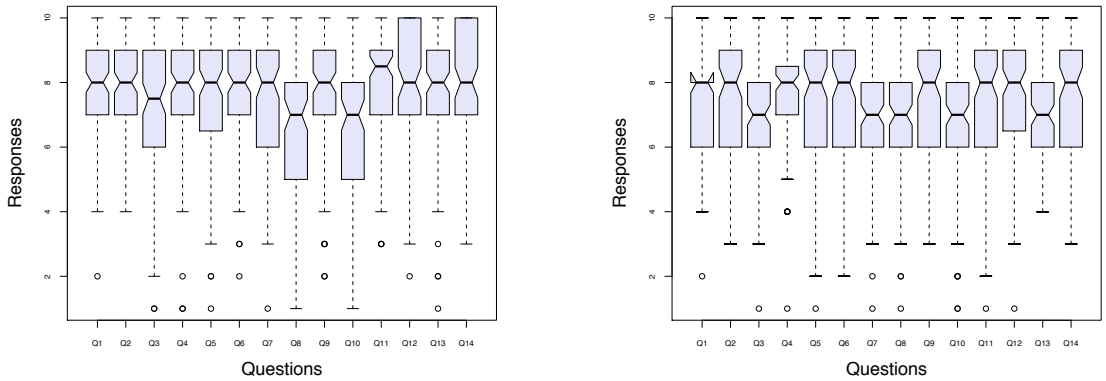

Figure 1: Boxplots of the responses for each questions from the $0.50 (left) and $0.05 (right) versions of the questionnaire.

by Kelly [8] over 23 students: 7.319 (sd=1.38; min=5.14; max=9.79).

4. EXPERIMENT PLATFORM

Our previous study of search frustration [6] was conducted in a lab setting, which allowed a variety of custom software and sensors to be deployed and logged participants' searching and browsing behavior in detail. Transitioning to an online experimental platform like AMT brings with it a cost in the richness of the information that can be logged, because the experimenter can no longer completely control and track the participant's environment.

For our current experiments, we will be using a platform designed for AMT that also retains some of the benefits of a lab study. In the lab, we log searching and browsing behavior using an open-source browser toolbar, the Lemur Query Log toolbar, which records not only queries and result click-throughs on the search engine, but also page views on target sites. Unfortunately, it is too much to expect AMT workers to install new software in their browsers, especially software that may inadvertently violate the worker's privacy in other browsing, unrelated to the assigned search task.

Instead, we log the search session by requiring the worker to issue searches and browse results through a proxy that we control. We show each Turker a page with a form and an imbedded frame, which points to the proxy. A diagram of the setup is shown in Figure 2. The form, located in the task pane, consists of the task to be completed and a text area where the user is required to respond to the task. The proxy frame is directed to a modified search engine interface made for the study. The proxy rewrites all links on every page that passes through so that those pages are redirected via the proxy as well. It injects JavaScript calls so that events, such as pages visited and mouse movements, can be logged. When the user has completed a task, they click the "Next" button in the task pane. This causes several hidden fields in the form to be populated with the events logged by the injected JavaScript. This data can either be uploaded to a database or sent to the outer frame in the AMT HIT page.

Alternatively, the proxy server could record a search log as pages pass through it, associating the log with a session identifier. We chose the JavaScript injection approach instead because it allows us to capture client side information, such as mouse movements. One could inject an off-the-shelf analytics package like userfly[1], which generates videos of browsing sessions, but we feel it would be more valuable to store low-level events directly in the search log for further analysis.

We have tested a variation of this platform by posting dozens of simple search tasks on AMT (such as "What is the record for the fastest mile run?" and "Who is the president of Harvard University?"), with a proxy frame included in the task, and successfully captured search logs from users on a variety of browsers. Two preliminary observations can be made, relevant to running these kinds of experiments on AMT. First, AMT workers copy-and-paste heavily, in order to work as efficiently as possible. As a result, the first query in many logs is the exact wording of the question, copied directly from the task frame into the search box. In experiments, it may be desirable to inhibit direct copying by presenting the task as an image, rather than as text. Second, a few workers answered the task without generating any search log, suggesting either that they already knew the answer or that they searched for it outside the proxy frame (contrary to the instructions of the task). This problem could be addressed by requiring use of the proxy frame before the answer can be submitted – e.g., by requiring that some part of the answer be selected, copied, and pasted from the proxy frame, which can be observed by selection events.

5. OPEN DESIGN ISSUES

We are currently in the process of completing the design for our search self-efficacy studies, and while we have determined there is much to be gained from using AMT to conduct this study, we have some open design issues to address.

5.1 Screening

We would like to include web searchers with both low and high self-efficacy in our study. AMT has the advantage of allowing a very large potential pool of study participants. We can administer the search self-efficacy scale as a screening tool, then administer our search assistance experiment to a stratified sample of users at different levels of search self-efficacy, ensuring that we screen sufficient numbers of users

[1] `http://www.userfly.com`

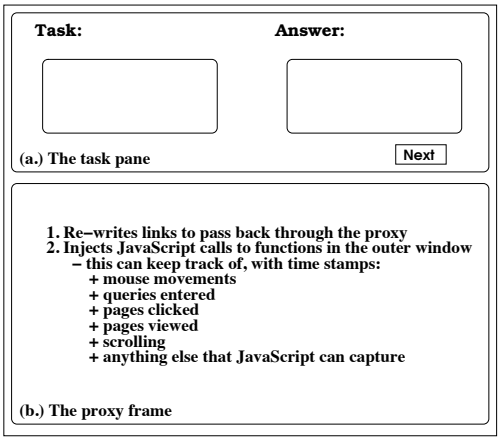

Figure 2: The proxy setup for the study.

to find enough at low self-efficacy levels. However, from our initial investigations, the range of search self-efficacy seen in Turkers appears to be similar to those of the undergraduate population [8].

5.2 Time to Completion

Time to completion raises questions of several kinds. One is consideration of the hourly wage being paid. An important ethical question we may want to examine is how much people are being paid for their work [9]. If the median time to complete a questionnaire is used as a marker, the hourly wage paid for the two HITs issued for this study were $15.31 and $1.95 per hour, respectively. The latter was an alarming number, and so we paid a bonus of $0.17 to each worker to increase the median rate to $8.55 per hour, the highest minimum wage in the United States as reported by the U.S. Department of Labor[2].

The second concern with time to completion is how much attention users give to their answers. For example, is the Turker answering the self-efficacy part of the questionnaire in 28 seconds actually reading the questions, or just filling it in arbitrarily as quickly as possible. This point is related to "instrument reliability", our next design issue.

5.3 Instrument Reliability

The questionnaire we presented to the Turkers contained only questions for which a high score means high self-efficacy. This does not provide us with any error checking. As mentioned in Section 4, Turkers have shown a tendency to complete tasks as efficiently as possible, which may include minimizing mouse movements. This means that one way to complete the questionnaire is simply to select items by location on the screen. We intend to experiment with different question wording in our full study. One technique that potentially enables identification of people who may not be engaging in depth with questions on surveys has been to provide questions with both positively and negatively phrased versions [2]. In our own work, this approach has allowed us to identify and filter out survey respondents whose answer profiles suggest they are selecting options so as to optimize time-to-completion and are therefore unlikely to be providing useful data [4].

[2]http://www.dol.gov/whd/minwage/america.htm

5.4 Truthfulness

While self-efficacy relates only to self-perception and not performance, we would like users' honest opinions about their self-efficacy. If Turkers view the questionnaire as a screening mechanism akin to a job interview, they may be incentivized to report higher self-efficacy than they truly feel. This is clearly an issue with all surveys of this kind, where participants often has a sense of what are "desirable" responses [2]. One way for us to address this is again through a slightly different phrasing on the questions such that desirable responses are not so clearly implied by the context (e.g., search ability is clearly a good skill to have and strongly aligned with being online—so Turkers likely skew towards seeing search prowess as desirable).

6. NEXT STEPS

The work described here gives us the ingredients needed for our full study. Our next steps are to complete the study using the following design:

- Modify the search self-efficacy scale so we can estimate reliability
- Screen Turkers with cross-checked search self-efficacy assessments to create a stratified sample by search self-efficacy
- Integrate search assistance mechanisms into the Turker search logging platform
- Design a post-survey about the level of task difficulty
- Evaluate the effects of search assistance, taking into account searcher self-efficacy and task difficulty

7. REFERENCES

[1] P. G. Anick. Using terminological feedback for web search refinement: a log-based study. In *SIGIR*, pages 88–95. ACM, 2003.

[2] A. Bandura. Self-efficacy: Toward a unifying theory of behavioral change. *Psych. Review*, 84(2):191–215, 1977.

[3] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proc. of WWW*, pages 1–10, 2009.

[4] C. Cheshire, J. Antin, and E. Churchill. Behaviors, adverse events and dispositions: An empirical study of online discretion and information control. *JASIST*, 61(7):1487–1501, 2010.

[5] D. Compeau and C. Higgins. Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly*, 189(211), June 1995.

[6] H. Feild, R. Jones, and J. Allan. Predicting searcher frustration. In *Proc. of SIGIR*, 2010.

[7] A. Hassan, R. Jones, and K. L. Klinkner. Beyond DCG: user behavior as a predictor of a successful search. In *Proc. of WSDM*, pages 221–230, 2010.

[8] D. Kelly. A preliminary investigation of search self-efficacy. Technical Report TR-2010-01, U. of North Caroline School of Information and Library Science, 2010.

[9] K. Mieszkowski. "I make $1.45 a week and I love it". *Salon.com*, 2006.