Studying User Browsing Behavior Through Gamified Search Tasks

Jiyin He
CWI
Science Park 123
1098XG Amsterdam
j.he@cwi.nl

Marc Bron
University of Amsterdam
Science Park 904
1098XH Amsterdam
m.m.bron@uva.nl

Leif Azzopardi
University of Glasgow
18 Lilybank Gardens Glasgow
leifos@acm.org

Arjen de Vries
CWI
Science Park 123
1098XG Amsterdam
arjen@acm.org

ABSTRACT

Typical crowdsourcing tasks ask workers to label images or make relevance judgements, as a low cost alternative to lab based user studies. More recently, gamification has been employed as a way to make these tasks more appealing and so users play, rather than work. One observation is that differences in task design and incentives elicits different player behavior. In this paper we discuss a new type of task, where we aim at eliciting player behavior that resembles user behavior when performing a search task. Care should be taken in the design of a gamified version of such a task to allow players to complete tasks with a limited amount of effort and time, without changing the behavior to be studied. We discuss the motivation of the abstractions and design choices we have made in achieving this goal. We then analyze whether and how these abstractions and design choices influence our observations of player behaviors.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.5.2 [**User Interfaces**]: Evaluation/methodology

Keywords

1. INTRODUCTION

Crowdsourcing as a collective problem solving approach has many applications. To increase user engagement, and to subsequently improve the quantity and quality of the crowd sourced tasks, often game elements such as a high score boards, badges, and progress bars are introduced [3].

Researchers have successfully applied gamification for different types of task. The main type of tasks involves asking workers to provide "objective" answers for labelling or classification tasks, e.g., to recognize whether there is a certain object in a video [5], or more specialized, e.g., to label fish species [4]. A more subjective tasks involves ascertaining opinions from the crowd, e.g., relevance judgements, which relies on consensus over subjective answers [2]. The task we discuss in this paper, however, is yet another type, which involves workers emulating users engaged in a search task with a particular search system. Such a task allows researchers to collect behavioral data to set the parameters of user interaction models with a particular search system. These interaction models have recently become a popular way to evaluate search systems.

Compared to a lab user study, a crowd-sourcing based study is likely to collect larger amounts of observations within a smaller amount of time and with lower cost. However, a difficult or lengthy task, such as a search task, may not attract any workers or they may quickly turn away for other easier, more interesting, or more lucrative tasks [6]. To crowd-source this type of task, it is necessary to simplify the task so that they can be accomplished within a reasonable amount of time and effort [8]. A key challenge is therefore to design the task in such a way that it allows the study of the behavior of interest while abstracting away from other possible behaviors.

In this paper, we present a case study where we study user browsing behavior with a faceted interface in a gamified setting. We describe the task design and abstractions made. We discuss our motivation for these design choices, and the impact of the task design and abstractions as observed from the data collected and the feedback from the players.

2. HOW USERS BROWSE (IN A GAME)

In this section, we introduce the background of the study and the design of search tasks and interfaces.

2.1 Study background

The development of a gamified task to study user behavior is part of a study in which we develop and validate a novel user interaction model that characterizes user interaction with a faceted search interface.

The user interaction model predicts the effort a user needs, in terms of the number of actions, to accomplish a search task, i.e., finding x relevant documents. Actions include examining a document to determine its relevance, pagination, and choosing a facet to filter results. To compute the predicted user effort, two model parameters need to be set: (1) a persistence parameter which describes how likely a user will continue to examine a next document at rank r – presumably a user is less likely to examine another document when he/she is already at the bottom of a ranked list, compared to when he/she is at the top of a ranked list; and (2) a facet selection parameter, which controls the likelihood a particular facet value is used to filter the results for a query.

Our primary goal in deploying the gamified task is to collect three quantities: the value of the persistence parameter, the facet selection parameter value, and the number of actions required to complete the task. Using these derived parameter values, we aim to compute the predicted user effort with our model, and compare

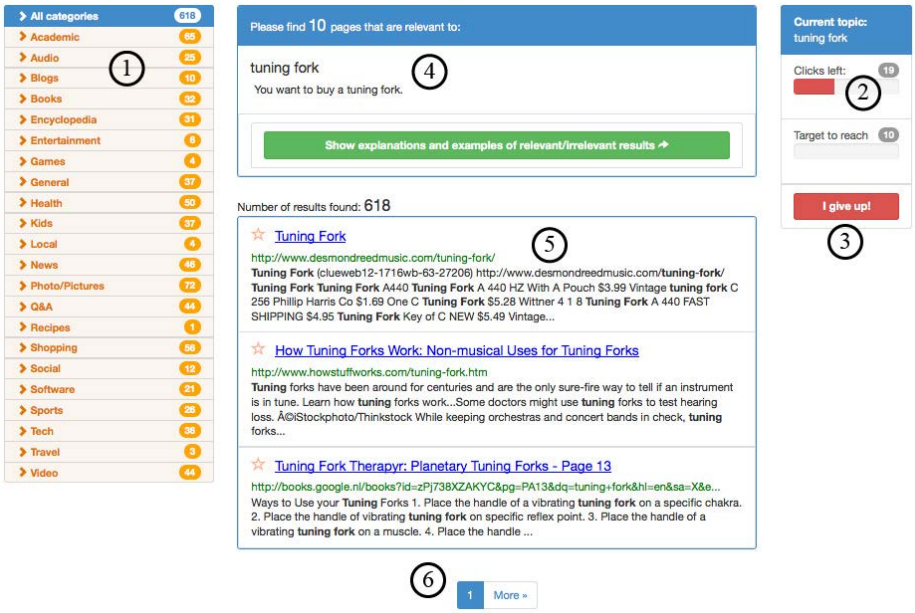

Figure 1: The faceted interface with left the facets (1), right the scoreboard (2) and give up button (3). The topic description (4) is available at the top. The middle of the screen is devoted to a scrollable result list with 10 snippets (5). At the bottom a pagination button (6) is available. The basic interface follows this design with the exception of the facet panel.

the result to the actual user effort collected in the game. This way, we are able to validate the accuracy of the model predictions.

A good alternative to crowdsourcing to obtain our parameter settings would be a lab study. Our motivation of conducting this study in a gamified setting instead of a lab user study is threefold. First, lab studies are expensive to carry out which limits the number of participants and tasks that can be included in the study. Second, participants may feel little engagement with a search task that asks to find information for which they have no actual need. Third, participants may require some time before they familiarize themselves with a novel interface and data collection. We discuss these points further after we introduce the gamified setting.

2.2 The Game Loop

We spread the link to the game through University mailing lists and social media. Clicking the link directs a participant to a login screen explaining the goal of the study. After creating an account a participant is asked to agree to a consent form explaining that all actions with the system are logged, that data collected during the study will be treated confidentially, and that it is possible to quit the study at any time. This is followed by a brief pre-questionnaire to collect basic demographic information. Next, a participant is presented with an instruction page on how to operate the search system interface and details on the how to complete a task. These instructions remain available throughout the game. After reading the instructions, a participant is presented with the topic description of the first task. After completing a task a participant is directed to a summary screen presenting his/her score, a leader board presenting the current scores of other participants and the option to complete another task. Clicking on the latter button, will direct a participant to the topic description for the next task.

2.3 Interfaces

Participants are presented with one of two interfaces: a basic interface, that mimics a standard search engine with a search box and 10 results, and a faceted interface that additionally provides a panel with facets. Figure 1 shows a screenshot of the faceted interface, where we use numbers, i.e., 1, ..., 6, to indicate various components of the system. The facets (1) are located on the left. Clicking on a facet removes all documents from the results that are not covered by that facet. At the top a topic description (4) is available. This is the description of the information need for which a participant is asked to find relevant documents. An additional button allows participants to expand the description and review the examples as provided before starting a task. The topics, facets, and relevance judgements are taken from the 2013 Federated Search track [1]. On the right a scoreboard (2) is available that provides participants with an overview of the number of clicks left and number of relevant documents found. After 25 clicks a give up (3) button appears providing the option to skip the remainder of the task. A scrollable result list with 10 snippets (5) takes up the middle part of the screen. Clicking on a snippet provides the user with feedback whether the document was relevant or not. We do not provide the ability to view documents. At the bottom of the page a pagination button (6) is available.

The design of the basic interface is identical to that of the faceted one with the exception of the facets, which are not available. Note that one of the facets provided in the faceted interface is the "all category" which is the unfiltered result list as it would be in the basic interface. This allows participants to ignore the facets and use the faceted interface as if they were using the basic interface.

2.4 Task design

Scoring rules. Participants were asked to find 10 relevant documents within 50 clicks. The following events invoke a click: clicking on a document (i.e., to indicate the selection of a relevant document), pagination, and clicking on a facet to filter results. After finding 10 relevant documents, the remaining clicks are added to the score of the participant, and shown in the score board.

The motivation behind this scoring rule is to make participants

think about their decisions. In order to achieve a high score, it is necessary to carefully examine the documents to determine whether it is likely to be relevant, as clicking on an irrelevant document would simply waste the available clicks. Further, since filtering and pagination also cost clicks, participants need to make decisions between continuing on the same ranked list or applying a filter.

To give up. Participants are able to give up after 25 clicks. Here, we seek a balance between the engagement of participants and the difficulty of the tasks. During pilot testing we found that immediately providing a give up button allowed participants to cherry pick easy tasks, while being trapped in a difficult task resulted in participants spam clicking to be able to move to the next task.

3. OBSERVATIONS AND FEEDBACK FROM PLAYERS

We collected usage data from 118 players and a total of 363 search sessions within a week after launching the game. After that, we closed the site as few new participants registered for the game or continued to play. Below, we discuss some of the observations made from the collected data, as well as from user feedback, with respect to the design choices we have made.

Relevance feedback. Normally, when searching, users formulate their own information need. They scan snippets in a result list and click on the link when a snippet seems potentially relevant. After examining the document the user would know whether it is relevant or not. In user studies participants are usually asked to solve a search task for an information need provided by the experimenters. As the information need does not originate from the participant he/she may have difficulties determining which documents are relevant (according to TREC assessors).

In the game, clicking on a result snippet results in the snippet turning green (with a check mark icon) or red (with a corresponding cross icon), indicating a relevant, or non-relevant document, respectively. This type of feedback allows us to mimic the process users go through in a real search task, i.e., that when they open (and read) a document they are able to determine whether it is relevant or not. The alternative is to allow users to make their own relevance judgements. However, the goal is to study the behavior of a user with the particular provided information need and associated relevance judgements not the possible variations in the information need.

Rank bias. Studies of user behavior on web search result pages, have observed a bias towards inspecting only documents at the top of the ranking [7]. To determine which summaries a user has visited we consider mouse hovers over results, which is shown to correlate with eye-gaze [7]. Figure 2 shows the percentage of total hovers over the 10 ranks of each result page. For the basic interface (black bar) we observe a difference of 3% between the highest and lowest rank. This difference is 6% for the faceted interface The distribution of hovers over the ranks is relatively uniform, i.e., compared to the distribution of hovers over ranks on Web search engine result pages where differences of 38% between the highest and lowest rank are observed.

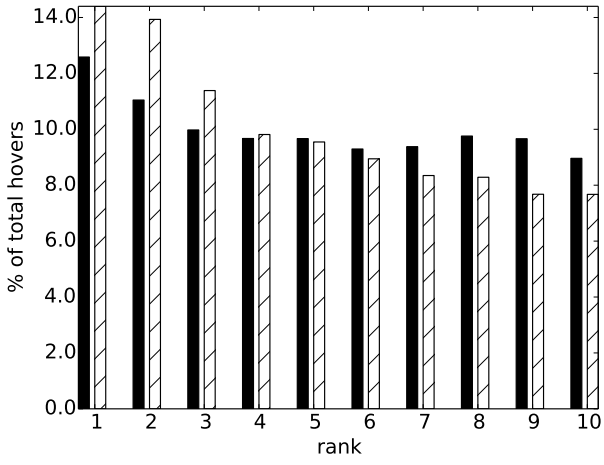

Figure 2: The distribution of total hovers over the 10 ranks of our result pages. The black bar indicates hovers observed with the basic interface. The white hatched bar indicates hovers observed with the faceted interface.

3.1 Feedback from players

A number of players commented on various aspects of the game through informal ways, e.g., on Facebook, in person, and via emails. We discuss these comments below.

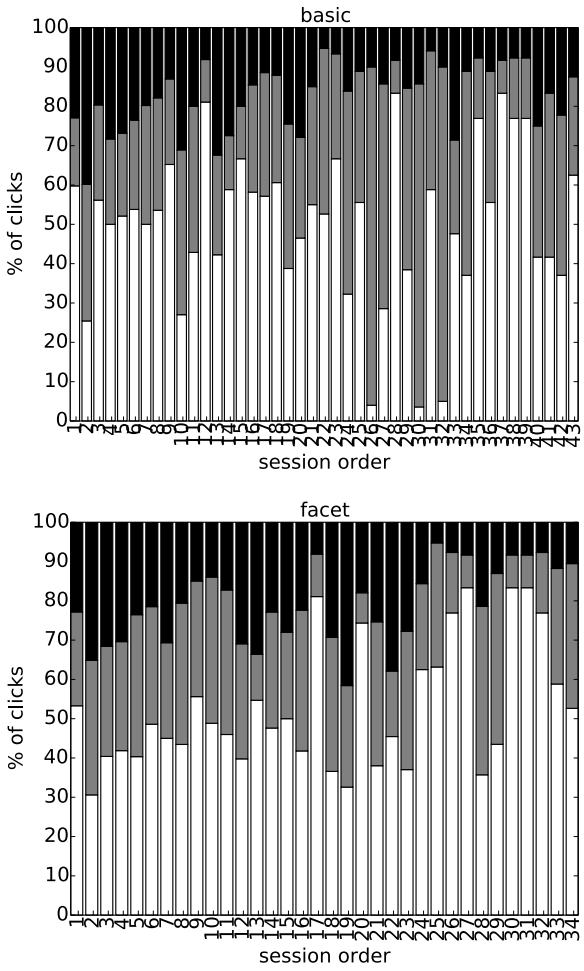

Figure 3: The percentage of clicks on relevant snippets (white), irrelevant snippets (gray), and other elements, i.e. facets and pagination (black) for the xth game session, aggregated over all users. On top the clicks with the basic interface, on bottom the clicks with the facet interface.

Relevance. Although we provided a description of the topic as well as examples of relevant and non-relevant documents, several participants voiced their concerns about the completeness and quality of the relevance judgements. Figure 3 shows the percentage of clicks on relevant snippets (white), irrelevant snippets (gray), and other elements, i.e. facets and pagination (black) for the xth game session, aggregated over all users. Note that the first game session is the first task a user completed, the second game session is the second task that user completed etc. As tasks are randomized user generally perform different tasks during their xth game session. We see that in all cases users click on non-relevant documents. Indeed, participants disagreed with the TREC relevance judgements, which are known to have a limit on the inter annotator agreement of about 70% [9].

Game elements. Some participants have commented that as far as games go, it is not particularly fun. We find that 145 tasks are completed by 49 participants using the basic interface while 255 tasks are completed by 48 participants with the faceted interface. Further, there are 35 uncompleted tasks (given up, failed or simply unfinished) with the basic interface, and 28 uncompleted tasks with the faceted interface. This suggests that the facets provide an additional challenge and play a role in user engagement.

Learning to use facets. Some participants commented that it took a while for them to figure out how to use the facets, e.g., "*it took me a while to figure out what the facets were for, but when I encountered a difficult query I found that they were useful and started to use them more*".

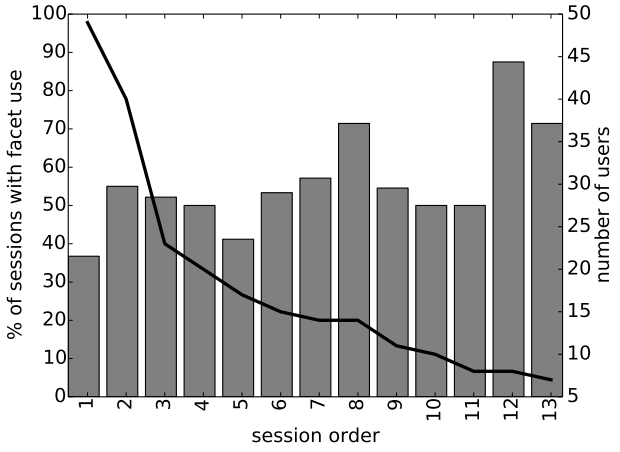

Figure 4: For each xth game session, aggregated over all users, the bars show the percentage of sessions in which facets are used.

To analyse this effect, we investigate whether and how the usage of facets changes as users play more sessions. Figure 4 shows the percentage of users that use facets at their xth game session. Here we only consider users that are assigned to a faceted interface. We note that few users complete more than 10 tasks. Although tasks are randomized and aggregated, in later sessions fewer tasks are available making the observation less reliable. We therefore only consider facet use up to the 13th session. Compared to later sessions, the first sessions have a relatively lower percentage in terms of facet usage. However, more than 30% of the users already started using facets in their first session. We observe the trend that facets are used more often in later sessions.

4. CONCLUSION

In this paper, we described a study that aimed at observing user browsing behavior through a gamified search task. We introduced the motivation and design choices made in order to simplify the search tasks, while striking a balance between abstraction of the task and allowing users enough freedom to exhibit natural behavior with the components under study. We find that in this setting participants find the faceted interface more interesting than the basic interface. Additionally, the rank bias normally observed in web search is not as prevalent, i.e., participants search further down the ranked list, which is most likely explained by the fact that users are asked to examine summaries instead of documents. Further, we discussed some of the participants' feedback on the TREC relevance judgements, which is consistent with what has been found in the literature. Finally, we find that in a gamified setting users quickly pick up on the use of facets.

We have described how users search with a gamified version of a search task. An interesting direction to explore is how the game rules can be adjusted to minimize the changes to the behavior to be studied. Further, a comparison to a non-gamified lab study will provide more insights into the influence of the design choices we have made on the observed user behaviour.

5. ACKNOWLEDGMENTS

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project nr 650.001.005.

6. REFERENCES

[1] T. Demeester, D. Trieschnigg, D. Nguyen, and D. Hiemstra. Overview of the trec 2013 federated web search track. In *TREC'14*. TREC, 2014.

[2] C. Eickhoff, C. G. Harris, A. P. de Vries, and P. Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *SIGIR'12*, pages 871–880. ACM, 2012.

[3] J. Hamari, J. Koivisto, and H. Sarsa. Does gamification work?-a literature review of empirical studies on gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences. HICSS*, 2014.

[4] J. He, J. van Ossenbruggen, and A. P. de Vries. Do you need experts in the crowd?: a case study in image annotation for marine biology. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 57–60, 2013.

[5] M. Hildebrand, M. Brinkerink, R. Gligorov, M. van Steenbergen, J. Huijkman, and J. Oomen. Waisda?: video labeling game. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 823–826. ACM, 2013.

[6] J. J. Horton and L. B. Chilton. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 209–218. ACM, 2010.

[7] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *SIGCHI'11*, pages 1225–1234. ACM, 2011.

[8] L. Von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.

[9] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36(5):697–716, 2000.