Investigating Factors Influencing Crowdsourcing Tasks with High Imaginative Load

Raynor Vliegendhart
R.Vliegendhart@tudelft.nl

Martha Larson
M.A.Larson@tudelft.nl

Christoph Kofler
C.Kofler@tudelft.nl

Carsten Eickhoff
C.Eickhoff@tudelft.nl

Johan Pouwelse
J.A.Pouwelse@tudelft.nl

Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

ABSTRACT

This paper reports useful observations made during the design and test of a crowdsourcing task with a high "imaginative load", a term we introduce to designate a task that requires workers to answer questions from a hypothetical point of view that is beyond their daily experiences. We find that workers are able to deliver high quality responses to such HITs, but that it is important that the HIT title allows workers to formulate accurate expectations of the task. Also important is the inclusion of free-text justification questions that target specific items in a pattern that is not obviously predictable. These findings were supported by a small-scale experiment run on several crowdsourcing platforms.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.1.2 [**Models and Principles**]: User/Machine Systems—*human factors*

General Terms

Design, Human Factors, Measurement

Keywords

crowdsourcing, Mechanical Turk, user study, quality control

1. INTRODUCTION

Crowdsourcing platforms increase the ease and speed with which new search functionality can be evaluated from a user perspective. In this paper, we take a closer look at issues that arise when a search-related feature is to be evaluated, but has not yet been implemented in working form into the system. The system in question is a file-sharing system. The evaluation takes place as part of the design cycle and has the purpose of allowing us to decide which of several possible realizations of the feature will be most effective for users of the system.

During the course of designing and testing the evaluation task for the crowdsourcing platform, we realized that our task was rather different in an important respect from other, more conventional, tasks carried out by workers on crowdsourcing platforms. Specifically, we needed the workers to be able to project themselves into the role of a user of the file-sharing system and to provide feedback from the perspective of that role. The projection is necessary for two reasons, first, because the system feature that we are evaluating does not yet exist, and second, because our target group of users are general, mainstream Internet users for whom the mechanics of file sharing is rather a stretch beyond their daily online activities. In an initial exploratory phase, we noticed that there was something "special" about our task. Few workers were choosing to carry out the HITs that we published to the crowdsourcing platform, and the batch completion time was longer than was acceptable given the time constraints of our design and implementation process. Our aim was to increase the number of participants in our HIT and also the rate at which new workers took up our HIT without changing the HIT in such a way that would discourage projection or attract cheaters.

In this paper, we report on this investigations that we undertook in order to design a HIT that would achieve this aim. First, we carry out an exploratory analysis of several experimental HIT designs on Amazon's Mechanical Turk (MTurk) and formulate our findings as a series of observations. Then, we build on these observations, performing a small-scale experiment on several crowdsourcing platforms. The experiment tests two aspects of HIT design (title and free-text justifications) that we found helpful for encouraging workers to undertake projection. We refer to tasks such as our evaluation task that require workers to project beyond tangible reality and beyond their daily experience as "crowdsourcing tasks with high imaginative load". We choose the designation *imaginative load* since we see certain similarities with tasks with a high cognitive load (e.g., they take relatively long, cannot be easily routinized and are difficult to carry out in highly distracting surroundings), but have concluded it is not possible to conflate such tasks with high cognitive load tasks, which would typically require using memory or at least some factual recall effort.

The contribution of this paper is a compilation of considerations that should be taken into account when using crowdsourcing for tasks with a high imaginative load, including suggestions for choices concerning HIT design and crowdsourcing platform that make it easier to design effective HITs for such tasks. Notice that we do not report the results of the evaluation itself in this paper. Rather, we concentrate on conveying to readers the information that

WSDM 2011 Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011), Hong Kong, China, Feb. 9, 2011

we acquired during the design of the evaluation tasks that we anticipate will be helpful in design of further tasks.

The paper is organized as follows. In the next section, we discuss related work (Section 2), then we describe the evaluation task (Section 3). In Section 4, we summarize our observations during the design and test of the task. In Section 5, we report on experiments carried out to investigate the impact of the titles and the verification on the behavior of the workers carrying out our HITs. Finally, in Section 6 we offer a summary of our conclusions.

2. RELATED WORK

In this section, we provide a brief overview of crowdsouring literature using techniques similar to ours. Often a crowdsourcing task will use a qualifying HIT to identify a set of workers who are suited to carry out the main task. In [1, 5, 4], recruitment and screening HITs were used to differentiate between serious workers and cheaters. In [3], methods to prevent workers from taking cognitive shortcuts are investigated. Many more workers completed the qualification HIT than returned to complete an actual HIT, an effect we also observe. Senstivity of workers to titles is mentioned in [5], who notice, as we do, that the selection of HIT titles influence their attraction to workers. In [2], experiments were carried out with different titles, pay rates, whether a bonus should be granted, and if so, whether this fact should be communicated to workers or not. Following the evaluation results, workers gravitate towards HITs with "attractive titles", i.e., titles which are easier to understand. In contrast to HITs that explicitly offer an additional bonus, easier-to-understand titles do not imply a high accuracy per worker. Free-text and open-ended response possibilities are often used to check whether workers had an understanding of the task, as in [5]. We make use of a similar approach, in particular asking for justifications of answers. In this respect, our work is related to that of [4], who conducted a subjective study about political opinions by asking workers to justify their given answers in free-text explanations. Giving an opinion often requires a certain degree of projection, which we equate with imaginative load. Note, however, that our task goes beyond asking mere opinions to asking workers to formulate an opinion about a feature that does not yet exist in a use context that is unfamiliar from their daily experience.

3. EVALUATION TASK

Our evaluation task involved assessing the usefulness of a time-evolving term cloud intended to make it possible for users to gain an understanding of the kinds of content that are available within a specific file-sharing system in order to facilitate browsing and search. The term cloud will offer users the possibility to find items within the system, but most importantly it is meant to allow new users unfamiliar with the system to quickly build a mental picture of what kind of content is available via the system. Users should not have to spend extensive time interacting with the system or trying out queries that are frustrating since they do not return results. In order to evaluate whether users have gained an understanding of the content available in the file-sharing system, we test their ability to distinguish five kinds of content available in the system (TV, music, books, movies and software) from five kinds not available in the system (current news, commercials, sports, how to videos and home videos). We compare this ability without the term cloud and with several different different cloud designs. Our HIT asks users to make a series of judgments on whether specific files exist in the file-sharing system. An example judgment is shown in Figure 1.

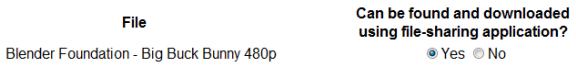

Figure 1: Example question from the evaluation HIT

Their answers to these questions will reflect whether or not users have generalized the information available in the term cloud into a mental picture that correctly represents the type of content in the system.

In order to prime workers to project themselves into the role of users of the file-sharing system and to discourage them from trying to use Internet search to determine which file-sharing system we are discussing and what sorts of files are present in it, we introduce the HIT with a "frame" that sets up an imaginary situation. The frame includes the following text and the diagram in Figure 2: *Jim and his large circle of friends have a huge collection of files that they are sharing with a very popular file-sharing program. The file-sharing program is a make-believe program. Please imagine that it looks something like this sketch:*

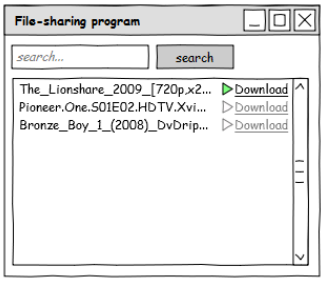

Figure 2: Mockup of a file-sharing program used to introduce (i.e., to "frame") our evaluation HIT

By naming a specific user of the file-sharing system, "Jim", we hope that users will better identify with a user of the file-sharing system, i.e., project themselves into that role.

We then ask for 10 worker judgments like the one in Figure 1. The HIT concludes with three validation questions, i.e., questions that do not ask for information necessary for the task, but rather allow us to judge the way in which the worker is approaching the task and eliminate low quality answers: (1) PrefQ, a personal preference question (multiple choice) *If you could download one of these files, which one would it be?* (2) PrefEx, a request to explain the personal preference (free-text question) *Why would you choose this particular file for download and viewing?* and (3) AnsEx, a request to justify one of the choices made while answering the 10 evaluation questions (free-text question) *Think again about the file that you chose. Why did you guess that Jim or one of his friends would have this file in their collection?* Note that there is an important difference between PrefEx, which asks workers to give a motivation for their

own opinion, and AnsEx, which asks workers to give a motivation from the perspective of the role in which we would like them to project themselves, i.e., a user of the file-sharing system.

We use multiple versions of this HIT, called the "evaluation HIT", in order to collect the information necessary for our study. Most of the cases discussed here are versions of the HIT that do not contain term clouds. We are interested in gauging the user's baseline evaluation answers before exposure to the term cloud. In some cases, we also use a recruitment HIT that establishes a closed pool of qualified workers. In the next section, we discuss observations concerning our HIT made during the design and test process.

4. EXPLORATORY ANALYSIS

This section provides a qualitative discussion of the issues that we encountered during the design and testing process of our evaluation. We relate these issues to the particular nature of our task—its high imaginative load.

Recruitment and worker volume Because the evaluation needed to fit our design and implementation schedule, it was important that our evaluation HITs quickly attracted an adequate volume of workers so that the total number of assignments associated with that HIT (i.e., the batch) completed within reasonable time. We soon noticed that workers from the recruitment HIT did not continue immediately on to carry out an evaluation HIT. We started our first evaluation HIT right after manually handing out qualifications to the 81 workers that completed our recruitment HIT successfully. Since the recruitment HIT took less than 24 hours to complete, we initially assumed that the evaluation HIT would complete within roughly the same amount of time. However, only 10 out of 405 HIT-assignments offered were completed the next day. A second recruitment yielded 79 new qualified workers, but only one of them took up the main evaluation HIT within 24 hours. We conjectured that this slow uptake was due to the mismatch in expectations raised by the recruitment HIT. The recruitment HIT was titled "Like movies and music? Earn qualification with a background survey and two short questions". It contained a list of relatively easy to answer background questions, but only one question containing titles as in Figure 1. In short, it did not reflect the focus of the main HIT. Workers were possibly misled to believe the main HIT would be more related to music and movies and did not expect to receive questions like Figure 1 in the main HIT. There are two possible interfering factors affecting the volume of workers: reward level, which we did our best to optimize before publishing this HIT, and total number of assignments available to workers. During a previous crowdsourcing project, e-mails from workers suggested that HIT popularity is related to offering a large volume of assignments and keeping them in steady supply. Because our recruitment HIT asks for free-text answers that must be individually judged, it is not possible to automate the assignment of qualifications in our evaluation, and for this reason the slow worker uptake was a real concern. We decided to publish an "open" evaluation HIT, i.e., one that did not require workers to earn a qualification, and were surprised that the quality of the responses to the free-text validation questions remained very stable. Apparently, our HIT has an aspect of its design that discourages workers who are not serious and makes recruitment less necessary.

Matching strategies. Because our evaluation task is attempting to gather information about people's mental pictures and not about the external world, there are no "correct answers" to the task questions. We could enlarge our HIT with questions for which the answer is known – a popular method for quality control – but the workers' ability to answer the control questions is not guaranteed to reflect the quality of their evaluation answers. For our task, it is more important to control for the strategy the worker is using to answer the question. In particular, we need the workers to be projecting themselves into the role of the user of the file-sharing application and not applying a strategy that reflects an external source of information (such as making use of general Internet search). A particular danger in the case of the evaluation HIT is that workers will try to apply a matching strategy using the information given in the "frame" of the HIT. In other words, it is possible that workers answer the evaluation questions by literally comparing the filenames in the example in Figure 2 or the terms in the term cloud (described in Section 3, but not pictured) to the filenames in Figure 1. Reading the explanations of why the workers thought that certain files were in the file-sharing system (i.e., the answer to AnsEx), it was clear that a few of the workers would base their decision on literal matches (e.g., one answers "cloud contains DVDRIP"). However, the majority were attempting to generalize the situation and make a decision on the basis of what kind of media enjoy overall popularity (in the case which does not include the term cloud) or what general categories of content are represented in the term cloud (e.g., one answers, "With the cloud screens showing words like programming and microsoft, I think this file should be available in the collection").

5. FURTHER INVESTIGATION

We carried out a small-scale experiment run on several crowdsourcing platforms in order to further investigate the impact of title choice and of the validation questions on the quality of the workers' responses. Each version of the HIT was made available to workers with a total of 50 assignments (5 sets of 10 different filenames to be judged) paying US$0.10 each. Results are reported in Table 1 in terms of batch statistics: number of assignments that we rejected due to obvious non-serious workers (e.g., blank text boxes), total number of workers participating, effective hourly rate, run time needed to complete the batch and median time between arrivals of new workers to work on the HIT-assignments.

Table 1: Batch statistics for the five experimental conditions (varying title and validation questions) on MTurk

	Title A	Title B	Title C	Only AnsEx	No PrefEx
#Rejected assignments	0	0	2	0	0
#Workers	25	22	19	17	20
Effective hour. rate	$2.54	$2.08	$1.76	$3.13	$1.51
Run time	50h28m	13h45m	19h13m	15h55m	20h45m
Med. arrival interval	67m36s	24m05s	18m41s	17m22s	35m06s

We experimented with three titles. Title A ("Jim, his friends and a make-believe file-sharing program"), which em-

phasized the imaginative nature of our HIT by including reference to "make believe". Title B ("Jim, his friends and digital stuff to download"), de-emphasized the fact that the HIT involved file sharing, terminology we thought might seem overly technical to workers. Title C ("Jim, his friends and interesting stuff to download"), which attempted to make the HIT generally attractive to a wide audience. We also experimented with omitting our validation question in order to understand which ones were important for maintaining high quality answers. We ran a version of our HIT which only asked for an explanation of the answer to the evaluation questions ("Only AnsEx") as well as a HIT that asked for a personal preference, but did not ask for that personal preference to be justified ("No PrefEx"). For completeness, we include a list of the limitations of this experiment, necessarily imposed by its small scale and short duration: We were able to control for temporal variation by starting each version of the HIT at approximately the same time on consecutive weekdays. We did not control for differences among weekdays or for the effects of holidays (for example, Title A ran the day before the Thanksgiving holiday in the US and we are careful not to read too much into its significantly longer runtime). We did not control for workers becoming acclimated to us as a requester and thereby more inclined to do our HITs. We simply checked that the number of workers that participated in multiple conditions remained limited (2–5). In this way, we know that our results are not dominated by workers who are developing strategies on how to approach the task from one HIT version to the next.

The following generalizations emerge from our investigation. First, all HITs yielded serious results—in only two cases did we reject an assignment completed by a worker due to blatant cheating. Second, the generally attractive title (Title C) seemed to attract workers at a better rate, but needed a longer total run time than Title B. Only requiring an explanation of the answer and not of personal opinion attracted workers quickly and also improved the total running time. However, here we noticed that we attracted two types of workers: first, workers who were taking the HIT seriously, spending relatively long to complete it and giving thoughtful answers to AnsEx and, second, workers who approached AnsEx with a "quick and dirty" strategy. Either these workers realized that the same answer was more or less applicable to all 5 sets of ten filenames and copied and pasted the same answer for each HIT-assignment that they completed or they fell into trivial non-specific observations, such as "That's what people share". In order to understand this effect, it is important to note that the wording of AnsEx was necessarily affected by the removal of the personal preference question from the "Only AnsEx" condition. It was no longer possible to ask for an explanation concerning the file that the user had picked. Instead of the original wording, the question was changed to "Think about the files that you thought were available for download. Why did you guess Jim and his friends would have these files in their collection?" This relatively small change meant that the question no longer targeted one specific file—the generality of the question apparently was enough to encourage non-serious workers to apply cut and paste strategies. Interestingly, the workers that answered the AnsEx question seriously in the "Only AnsEx" version of the HIT gave more elaborate answers than the workers doing the version of the HIT that required them to answer multiple validation questions. Also interesting was that the "No PrefEx" condition, which omitted the question requiring workers to justify their personal interests, yielded thoughtful answers on the AnsEx question, suggesting that the PrefEx question is not necessary. We would like to note that because the number of workers was relatively small, a single worker with a particular style (e.g., tending to apply a matching strategy) could have an inordinately large influence on the outcome of the experiment. If it is not possible to completely control for worker style, it appears important to use a quite large pool of workers in order to ensure the generality of results.

We ran the same set of experiments on other available crowdsourcing platforms to make a cross-platform comparison. Gambit and Give Work did not yield any judgments at all. This finding was largely independent of the financial reward offered. We conjecture that the lack of uptake may be due to technical limitations (mobile device, etc.) or a consequence of a different culture of HITs on these platforms. Samasource seems to be a very difficult platform to use. There were several negative observations to be made with our current experiment setup: Largely independent of title or question style we notice a very high share of uncreative copy and paste answers. Additionally there seem to be issues with their worker identification system as we have multiple submissions from different worker ids, that were issued from the same IP address and contained identical copy & paste answers. The very impressive exception to this trend was one worker from Nairobi who provided extremely detailed, informed and well-written answers.

6. CONCLUSIONS

We conclude that "high imaginative load" tasks can be successfully run on MTurk. The key appears to be a combination of signaling to workers the unique nature of the task, possibly quite different than tasks they generally choose, and at the same time making each HIT-assignment require a highly individualized free-text justification response.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission's 7th Framework Programme (FP7) under grant agreement № 216444 (NoE PetaMedia).

8. REFERENCES

[1] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor. Are your participants gaming the system? Screening Mechanical Turk workers. CHI '10, pages 2399–2402, 2010.

[2] C. Grady and M. Lease. Crowdsourcing document relevance assessment with Mechanical Turk. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's MTurk*, pages 172–179, 2010.

[3] A. Kapelner and D. Chandler. Preventing satisficing in online surveys: A "Kapcha" to ensure higher quality data. In *CrowdConf ACM Proceedings*, 2010.

[4] S. A. Munson and P. Resnick. Presenting diverse political opinions: how and how much. CHI '10, pages 1457–1466, 2010.

[5] K. T. Stolee and S. Elbaum. Exploring the use of crowdsourcing to support empirical studies in software engineering. ESEM '10, pages 35:1–35:4, 2010.