# Crowdsourcing Interactions

## A proposal for capturing user interactions through crowdsourcing

G. Zuccon, T. Leelanupab, S. Whiting, J. M. Jose, and L. Azzopardi
School of Computing Science, University of Glasgow
Glasgow, G12 8RZ, United Kingdom
{guido,kimm,stewh,jj,leif}@dcs.gla.ac.uk

## ABSTRACT

Crowdsourcing has been proposed as an inexpensive and often efficient way of outsourcing work tasks to a large group of people. In this paper, we propose to use crowdsourcing as strategy for acquiring users interactions within interactive information retrieval (IIR) systems. We are interested to understand whether crowdsourcing represents a robust strategy that can be used in conjunction with common approaches for capturing interactions, in particular laboratory user experiments. What are the similarities, differences, advantages and disadvantages of crowdsourcing interactions compared to traditional strategies? To investigate these issues, we outline the design of a procedure where interactions can be captured using crowdsourced workers. We expose the problematic issues that arise during the design process, together with preliminary statistics and results acquired by implementing our protocol within Amazon Mechanical Turk. This work opens up a number of research prospectives, the most appealing being a new methodology for the evaluation of IIR systems based on crowdsourcing.

## 1. INTRODUCTION

Collecting interactions between users and search system is fundamental for the analysis of user behaviours and system efficiency in the study of IIR systems. Users' interactions are often captured using search system logs. In practice, such interaction logs can be acquired in two ways: either from the system logs of search engines, in a completely naturalistic manner, or by set up experiments where users are invited to perform some pre-defined simulated information seeking tasks. Both techniques have several advantages, as well as disadvantages.

Obtaining search interactions through the analysis of query-logs generated by search engines whilst inexpensive, is virtually impossible, unless the organisation who owns the search engine grants access to this resource. Unfortunately, this is often not the case for academic researchers [5]. A further problem is that there is no control on the user population whose interactions have been captured. In fact, in these cases neither researchers have entry data about the user population, such as demographical information (e.g. age, sex, nationality, education, etc), level of confidence with the search technology and the information seeking task, nor can they obtain post-search task feedback from the users, such as their level of satisfaction about the search experience, level of achievement of the search goals, etc. These disadvantages are however mitigated by the availability of a large number of (often) heterogeneous user interactions, since everything users search for is logged by the retrieval systems.

Conversely, setting up laboratory user studies to capture search interactions with IIR systems is generally costly, as participants are usually paid at the minimum hourly wage. This limits the number of participants in laboratory-based user studies. Thus, the collected data is often several orders of magnitude smaller than what is acquired by search engines' query-logs. Moreover, participants are often recruited within an homogeneous user population. For example, in the case of researchers based within universities, users are often recruited within the university's student population. However, in laboratory user studies researchers have extensive control over the participants. Population observations such as demography, familiarity with search technologies/tasks, etc., can all be collected. Similarly, post search task feedback can be acquired explicitly from the users, e.g. using questionnaires or interviews.

In this paper, we propose an alternative approach to the IIR experiment methodology, based upon *crowdsourcing*. Crowdsourcing has been proposed as an inexpensive and often efficient way to conduct large-scale focused studies [7], and has been implemented in a number of web-based platforms such as Amazon Mechanical Turk (AMT) and CrowdFlower. The crowdsourcing paradigm has been recently used in information retrieval for performing a number of tasks. For example, Alonso et. al. crowdsourced relevance assessments by asking workers to evaluate the relevance of results retrieved by a geographical IR system [2]. While, Alonso and Mizzaro compared crowdsourced relevance judgements against the correspondent judgements obtained by TREC assessors [1] .

The intuition underlying the crowdsourcing-based user experiments we propose is that workers are asked to complete information seeking tasks within a web-based crowdsourcing platform. While workers perform information seeking tasks, researchers can capture logs of workers interactions with the IIR system. Furthermore, researchers have the possibility to acquire entry and post-search information and statistics, which would help to characterise (to some extent) the user population. This procedure might appear similar to laboratory-based experiments, and for this reason in this paper we focus on these two strategies. Note however that the inherent characteristics of crowdsourcing differentiates the two strategies. In section 3, we examine the diversities between crowdsourcing-based and laboratory-based (which


*CSDM'11,* WSDM 2011 Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011), Hong Kong, China, Feb. 9, 2011
ACM 978-1-60558-896-4/10/07.

are reviewed in section 2) IIR paradigms for capturing user interactions in information seeking tasks. The presence of such diversities calls for the definition of a new protocol for crowdsourcing-based IIR experiments, which is outlined in section 3. Thereafter we describe how we plan to validate the protocol, and we report preliminary results and statistics, together with open issues of the IIR experiments we executed on AMT (section 4). Finally, in section 5 we discuss the prospectives that the protocol for IIR experiments based on crowdsourcing opens up for IR research, the most appealing being a novel methodology for the *evaluation* of IIR systems.

# 2. TRADITIONAL PARADIGMS FOR INTERACTIVE IR EXPERIMENTS

The acquisition of interaction data in laboratory based IIR settings follows the indications put forward by Borlund in the paradigm for the evaluation of IIR systems [3]. Within this scheme, information needs are treated for individual users with respect to search tasks and simulated situations. In a cognitive perspective, the knowledge and the perception of the search context is represented by the user's interactions with the IIR system. According to Borlund, the IIR evaluation model is composed of three main aspects:

a. Experimental *settings and protocols*, e.g. the Latin-square procedure, and sequences of pre-defined tasks and questionnaires.
b. Simulated *work task* situation, that provides adequate imaginative context to initiate user's information needs.
c. Set of alternative *performance measures* regarding user behaviours via logs (e.g. completion time, entered queries terms, read results, etc.), as well as user perception and search experience via questionnaires and interview (e.g. open question, Likert scale, or semantic differentials).

In the traditional IIR model the focus of the evaluation is on the behaviour of users performing search. During the search session, a user interactively searches, interprets, and modifies the search as well as the relevance assessment with respect to the perception of information needs and simulated task situations. Furthermore, Borlund suggests that participants should be from backgrounds similar to the simulated situation designed by the experimenters. However, this is not always the case, as university students are often employed by academic researchers for performing laboratory based IIR experiments, e.g. [9].

# 3. A PROTOCOL FOR INTERACTIVE IR BASED ON CROWDSOURCING

In the following we outline a protocol for conducting IIR experiments, and thus capture interaction data, within crowdsourcing platforms. Some of the considerations we develop in the following are based on the tools provided by AMT, but can be extended and adapted to other crowdsourcing platforms, such as CrowdFlower.

The protocol we propose prescribes that workers are asked to perform self-contained information seeking tasks within a unit of work (also known as HIT in AMT) advertised on the crowdsourcing platform. In the meantime, researchers can collect logs of workers' interactions with the IIR system as well as post-search information and statistics. Although this procedure might appear similar to laboratory based IIR experiments, a number of key factors affect important experimental aspects, thus effectively differentiating these two strategies. In particular, they differ in the way the following experimental aspects are tackled:

1. Characterise user population (section 3.1)
2. Define information seeking tasks (section 3.2)
3. Capture interactions (section 3.3)
4. Acquire post-retrieval information (section 3.4)

Before describing the experimental aspects that characterise the protocol based on crowdsourcing, we briefly outline some of the key factors that differentiate crowdsourcing-based from laboratory-based IIR experiments.

**Heterogeneity.** The user population that can be reached through crowdsourcing is highly heterogeneous with respect to location, nationality, education, employment, age, sex, language, etc (see [8] for a demographical study of the workers of AMT).

**Cost.** Crowdsourced IIR experiments are likely to be cheaper than laboratory-based ones: e.g. the average hourly rate of the experiments detailed in section 4 is $1.38, while the national minimum wage in UK is about $9.35.

**Scale.** Because researchers can access a large number of workers through crowdsourcing tools, and because of the associated low costs, crowdsourcing often provides the opportunity to reach a higher number of participants for IIR experiments than laboratory-based approaches.

**Users' information quality.** While it is often assumed that participants in laboratory-based experiments provide to researchers correct and detailed information about themselves[1], the same cannot be assumed for crowdsourced workers. In fact, usage regulations of web-based crowdsourcing platforms often forbid researchers to ask for personal details of users (e.g. see AMT policies[2]). Furthermore, it cannot be excluded that malicious users participate in crowdsourced tasks. Finally, crowdsourced workers likely optimise their working strategy for completing tasks, so as to achieve task completion with the minimum effort or within a minimum time.

**Typology of IIR tasks.** In section 2 we have pointed out that traditional IIR experimental paradigms prescribe the creation of simulated work task situations. This often requires participants to read instruction sheets that not only outline how to use the IIR system, but also describe the simulated situation the user has to imagine and the information need he is expected to satisfy. This procedure is unlikely to be suitable for crowdsourced workers, as previous studies noted that the instructions provided to workers have to be kept short and simple, and workers are unlikely to perform the cognitive effort required by simulated situations and information seeking tasks.

**Quality of interactions/reliability of interactions.** Previous studies suggested that crowdsourced workers tend to complete tasks as efficiently as possible [6]. Furthermore, others suggested that malicious workers might submit tasks without actually performing the requested operations. These aspects pose doubts on the quality and reliability of interactions captured through crowdsourcing. Interactions obtained via crowdsourcing should be validated and then compared against those acquired with traditional approaches.

[1]Researchers select a group of qualified subjects and ask their personal information.

[2]`https://requester.mturk.com/mturk/help?helpPage=policies`

## 3.1 Characterise User Population

Pre-experiment questionnaires and interviews are usually employed by researchers for acquiring demographical and self-perceptual information about participants in laboratory-based IIR experiments. This method is however inapplicable for crowdsourced IIR experiments. First, if workers are asked to fill in questionnaires[3] within a unit of work (i.e. HIT), then they will have to enter the same information several times: as many as the number of HITs they perform. This problem can be overcome by requiring workers to pass a *qualification test*. By employing qualification tests, researchers can acquire background information about the users to characterise the user population. Furthermore, experimenters can exclude from their HITs those workers that do not meet pre-defined criteria suitable for the experiment. Once workers are characterised through a qualification test, they can be classified within groups on the basis of similar scores. Groups can then be used to compare and contrast search behaviours and interactions of crowdsourced workers against the ones obtained by correspondent groups of laboratory based participants. This approach provides a means for comparing search behaviours and interactions between the two user populations.

However, crowdsourcing tools do not usually allow requesters to ask personal questions to users, such as their age, sex, etc. Moreover, it is yet unclear how to judge the truthfulness of answers related to self-perception questions, such as workers' confidence with search engines and search tasks, their expertise, etc [6]. Thus, qualification tests have to be carefully chosen in order (i) not to violate the crowdsourcing tool's policies, (ii) to avoid doubts on the truthfulness of the acquired data, (iii) but yet to obtain information that characterises users and their abilities. To address these points, we propose to use qualification tests based on aptitude or Intelligence Quotient (IQ) tests developed in Psychometrics [4]. A further use of this test is to assess whether workers are suitable for the typology of information seeking tasks that are used in the experiments (e.g. domain specific applications). The intuition is that these tests provide a measure of reasoning skills, language knowledge and problem solving skills of crowdsourced workers, as well as a measure of their attention when performing crowdsourced tasks. It is yet to be said whether high IQ scores correspond to higher abilities in solving IIR tasks: this has to be further investigated. However, we expect that there is not a predominant score (or range of scores) amongst the ones obtained by crowdsourced workers. Conversely, we expect that if the same tests were performed by participants of laboratory studies recruited amongst the student population of universities, the scores would be predominantly grouped within a high score range, mainly because of the level of education of the participants, and for the fact that participants have often been already screened by universities[4] according to IQ tests when beginning their university degrees.

## 3.2 Define Information Seeking Tasks

Information seeking tasks assigned to crowdsourced workers have to be clear and well defined, as no interaction is possible between workers and requesters. Workers are unlikely to perform the cognitive effort required by simulated situations and information seeking tasks, as workers' main goal is to complete tasks as efficiently and rapidly as possible. We suggest that in crowdsourced IIR environments, researchers should explicitly provide the topic that the search will be about, together with a number of specific informational questions the workers are expected to answer. For example, one of the topics contained in the experiments we report in section 4 is "Australian wines". With respect to this topic, workers are asked to answer the following questions[5]: "What winery produces Yellowtail?", "Where does Australia rank in exports of wine?", and "Name some of Australia's female winemakers". We argue that posing questions about a specific topic initiates in the workers the search requirements needed by the settings of IIR experiments. We thus posit that no simulated tasks are required. In fact the scenario in which the information seeking task is performed results clear: workers have to answer a number of questions, and to help themselves they can find information about these questions by searching through the provided IIR system. Topics and questions should be carefully chosen so that answers are not likely to be known, and search needs are thus effectively initiated.

## 3.3 Capture Interactions

Once topics and questions are assigned, workers can search with the provided IIR system in order to find useful information for formulating answers. It is imperative for the IIR system to capture the interactions between the workers and the system itself (e.g. issued queries, clicked results, time spent in reading/searching, etc). Crowdsourcing platforms, such as AMT, do not provide native tools for capturing these kind of user interactions. However, several solutions can be devised so as to direct the workers towards a tool that is controlled by experimenters, and thus records workers' interactions. For example, Field et al. used a proxy to achieve this goal [6]. In the experiments reported in section 4 a different solution was adopted: workers were shown the interface of the IIR system within a self-contained iFrame positioned in the page of the HIT. Through iFrames, interactions could be recorded, making them available for further analysis.

## 3.4 Acquire Post-Search Information

Self-perception information about the search task workers just performed can be acquired by means of a questionnaire within a unit of work. Questions can be related to the difficulty of the task, the level of satisfaction with both system and answers provided, etc. However, little can be said about the truthfulness of the acquired data [6]. Nevertheless, this problematic issue can be partially addressed by well known techniques, e.g. different phrasing of subsequent questions, so that answers cannot be inferred by the context.

# 4. EXPERIMENTING WITH THE NEW IIR PROTOCOL

For the purpose of setting up a preliminary investigation of the novel protocol for IIR experiments introduced in section 3, we asked AMT's workers to carry out 24 search tasks[6] extracted from the TREC 2006 and 2007 Question-Answering track[7]. For each topic, three questions were se-

[3]We ignore the possibility of performing interviews of workers, given the remote and asymmetric nature of crowdsourcing.

[4]At least in many European countries.

[5]Of course, making use of a search engine we provide for helping them find information useful for answering the questions.

[6]Each task was repeated by three different workers.

[7]`http://trec.nist.gov/data/qamain.html`

lected. Twenty-three workers performed our HITs (a worker completed on average 3.13 HITs). Workers were divided into two groups: the first needed to pass a 20-questions aptitude qualification test, while the second did not. Workers completed units of work by answering the posed questions using the provided IIR system to assist them in finding the answers on the Web. They were also asked to mark Web pages containing information useful for answering the questions. After the questions were answered, workers were asked to evaluate various aspects of their search experience in a post-search questionnaire: 4 five-point semantic differential questions focused on the performed task, while 3 five-point Likert scales assessed their background knowledge of the task and the search they performed.

## 4.1 System Platform

We embedded our IIR system within the crowdsourcing platform offered by AMT, using an iFrame within a standard HIT. Our IIR system was developed as a web-based front-end of the Microsoft Bing API[8] for web results. Each time a user began a search task our system was provided with AMT HIT details such as the work assignment ID and the corresponding question topic. Queries, result clicks and explicit feedback via an optional "Mark as Relevant" button were logged alongside the HIT information. Following completion of the batch of HITs for each experiment we then merged the provided search logs with the AMT logs to yield a rich source of individual worker data for analysis. AMT data provided statistics such as the search task duration, question answers, unique worker IDs and qualification scores that can be used to begin explaining behaviours observed through the related query logs.

## 4.2 Preliminary Results

In table 1 we outline preliminary interaction statistics that were acquired through the experiments performed with the novel protocol for IIR based on crowdsourcing. A definitive statement deriving from the analysis of the reported data cannot be made yet, since at the moment we have not performed a laboratory-based counterpart of the experiment. However, the statistics show how workers had to interact with the search system in order to find information that helped them formulating answers to the provided questions. Moreover, feedback from workers show that the tasks we developed based on the TREC Question-Answering track were clear, slightly difficult, moderately complex, but familiar. Workers also stated that they did not know the answers before performing the HITs. In addition, they felt they successfully answered the questions.

## 4.3 Open Issues and Future Work

A number of issues have still to be investigated in order to assess the validity of the protocol we outlined in this paper:

1. Are the interactions acquired through crowdsourcing similar to those acquired through laboratory experiments? And, how can they be compared?
2. Is a training session required for crowdsourcing based experiments, as suggested by Borlund [3]?
3. Is it legitimate to use an aptitude test (IQ) to characterise and compare users in IIR settings?
4. What is the role of a crowdsourcing-based experiment in IIR evaluation? Can this be used to replace or complement laboratory-based experiments for qualitative and quantitative assessment?

[8] `http://www.bing.com/developers`

| | MIN | AVG | MAX | STD |
|---|---|---|---|---|
| Search Queries | 1.00 | 6.54 | 11.00 | 2.92 |
| Total Unique Viewed Pg. | 0.00 | 2.20 | 9.00 | 1.76 |
| Unique Viewed Pg. from Wiki | 0.00 | 0.45 | 3.00 | 0.73 |
| Unique Viewed Pg. from Non-Wiki | 0.00 | 1.75 | 7.00 | 1.64 |
| Total Unique Rel. Pg. | 0.00 | 0.73 | 4.00 | 1.17 |
| Unique Rel. Pg. from Wiki | 0.00 | 0.09 | 1.00 | 0.29 |
| Unique Rel. Pg. from Non-Wiki | 0.00 | 0.64 | 4.00 | 1.12 |
| Time Spent in Seconds | 36.00 | 459.86 | 776.00 | 195.68 |

**Table 1: Statistics of user interactions on 24 HITs.**

# 5. PROSPECTIVES FOR IIR

In this paper we have proposed a new strategy based on crowdsourcing for acquiring interactions between users and IIR systems. The acquisition of interaction data via crowdsourcing is not intended to act as a substitute in laboratory-based experiments, but complements it by offering additional data to analyse.

Moreover, if the validity of the proposed experimental protocol is confirmed by further studies, this work opens up a number of novel research prospectives for IIR. In fact, interaction data can be acquired following the proposed crowdsourcing protocol as to study querying behaviours, search strategies, and, ultimately, for comparing, contrasting and evaluating interactive IR systems.

Future work will be directed towards the consolidation and evaluation of the introduced crowdsourcing protocol for IIR, in particular by comparing the acquired information against that obtained through laboratory based experiments. Furthermore, we intend to explore the possibility of applying the protocol to the evaluation of IIR systems.

**Acknowledgements:** This work has been partially supported by the EPSRC Renaissance project (EP/F014384/1) and by the Royal Thai Government.

# 6. REFERENCES

[1] O. Alonso and S. Mizzaro. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *SIGIR 2009 Work. on The Future of IR Eval.*, 2009.

[2] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42:9–15, 2008.

[3] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003.

[4] P. J. Carter. *IQ and Psychometric Tests*. Kogan Page, 2007.

[5] W. B. Croft, M. Bendersky, H. Li, and G. Xu. Query representation and understanding workshop. *SIGIR Forum*, 44(2):48–53, 2010.

[6] H. Feild, R. Jones, R. C. Miller, R. Nayak, E. F. Churchill, and E. Velipasaoglu. Logging the Search Self-Efficacy of Amazon Mechanical Turkers. In *SIGIR 2009 Work. on Crowdsourcing for Search Eval.*, 2009.

[7] J. Howe. The Rise of Crowdsourcing. (accessed July 20, 2010), June 2006.

[8] J. Ross, A. Zaldivar, L. Irani, B. Tomlinson, and M. S. Silberman. Who are the crowdworkers? shifting demographics in mechanical turk. In *Proceedings CHI 2010*, pages 2863–2872, 2010.

[9] R. W. White. *Implicit Feedback for Interactive Information Retrieval*. PhD thesis, University of Glasgow, 2004.