# Impact of HIT Design on Crowdsourcing Relevance

Gabriella Kazai[1] Jaap Kamps[2] Marijn Koolen[2] Natasa Milic-Frayling[1]

[1] Microsoft Research, Cambridge UK
[2] University of Amsterdam, The Netherlands

## ABSTRACT

In this paper we investigate the design and implementation of effective crowdsourcing tasks in the context of book search evaluation. We observe the impact of aspects of the *Human Intelligence Task* (HIT) design on the quality of relevance labels provided by the crowd. We assess the output in terms of label agreement with a *gold standard* data set and observe the effect of the crowdsourced relevance judgments on the resulting system rankings. This enables us to observe the effect of crowdsourcing on the entire IR evaluation process. Using the test set and experimental runs from the INEX 2010 Book Track, we find that varying the HIT design and the pooling and document ordering strategies leads to considerable differences in agreement with the gold set labels. We then observe the impact of the crowdsourced relevance label sets on the relative system rankings using four IR performance metrics. System rankings based on MAP and Bpref remain less affected by different label sets while the Precision@10 and nDCG@10 lead to dramatically different system rankings, especially for labels acquired from HITs with weaker quality controls. Overall, we find that crowdsourcing can be an effective tool for the evaluation of IR systems, provided that care is taken when designing the HITs.

## 1. INTRODUCTION

The evaluation and tuning of Information Retrieval (IR) systems based on the Cranfield paradigm requires purpose-built test collections, at the heart of which lie the human relevance judgments. With the ever increasing size and diversity of both the document collections and the query sets, gathering relevance labels by editorial judges has become a challenge. Recently, *crowdsourcing* has emerged as a feasible approach to gathering relevance data. However, the use of crowdsourcing presents a radical departure from the controlled conditions in which editorial judgments are collected. In this paper, we explore the effectiveness of various HIT designs as a means of controlling the crowd workers' engagement and, consequently, the quality of the resulting relevance labels and the reliability of the IR evaluation in terms of the relative system rankings.

We focus our investigation of HIT designs on three aspects: 1) quality control elements, 2) document pooling and sampling for relevance judgments by the crowd, and 3) document ordering within a HIT for presentation to the workers. Based on the analysis of the collected data, we provide insights on 1) how design decisions influence both the raw label quality, i.e., agreement with *gold standard* (GS) obtained from traditional editorial judges, and 2) the usefulness of crowdsourced relevance labels in IR evaluation, i.e., their impact on the system rankings.

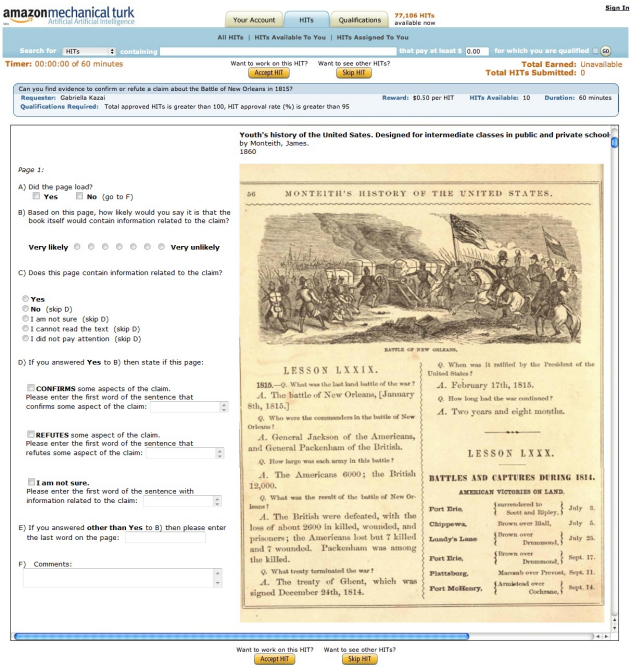



**Figure 1: Part of a HIT showing question series to solicit relevance labels for book pages from workers on Amazon Mechanical Turk: Full design.**

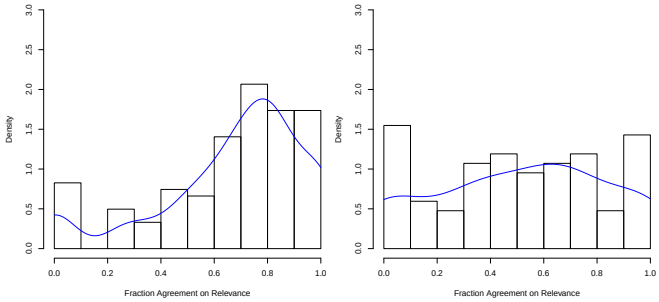
## 2. EXPERIMENTAL SETUP

**Data** We use the books, search topics, official runs, and relevance judgments provided by the INEX 2010 Book Track. The corpus comprises 50,239 out-of-copyright books, containing over 17 million pages and amounting to 400GB. There are 15 Best Books runs (Ad hoc retrieval of whole books) and 10 Prove It! runs (return pages that confirm or refute a claim).

**HIT designs** used two different sets of quality control mechanisms. *Full design (FullD)*, see Figure 1, controls all the stages of the task and explicitly pre-qualifies workers, re-

*This is an extended abstract of Kazai et al. [1].


*DIR-2012* February 23–24, 2012, Gent, Belgium.




(a) Full Design (b) Simple Design

**Figure 2: Distribution of workers over agreement as histogram and probability density function.**

**Table 1: System rank correlation between the different designs**

| Design | MAP | Bpref | P@10 | nDCG@10 |
|---|---|---|---|---|
| FullD | 0.76 | 0.45 | 0.85 | 0.73 |
| SimpleD | 0.96 | 0.87 | 0.34 | 0.02 |

stricting participation to those who completed over 100 HITs at a 95+% approval rate. It includes trap questions, qualification questions, a captcha, and dependencies between the questions. *Simple design (SimpleD)* does not impose restrictions on the workers who can participate. No qualifying test is included to check if workers are familiar with the claim. No warning is displayed to workers about the expected quality of their labels. Finally, no captcha is used in this design, also simplifying the structure of the Flow questions.

**Pooling strategies** We used three interleaved pools: 1) *Top-n pool* based on the submissions; 2) *Rank-boosted pool* reranking the submissions based on the *popularity* of books; 3) *Answer-boosted pool* reranking insisting on keywords of the topic being present on the page.

**Page ordering** We used two ways of dividing the pages over HITs of 10 pages: *Biased order* preserves the order of pages produced by a given pooling approach, so with decreasing expected relevance, and inserts a known relevant page at the first position in the HIT. *Random order* first inserts a known relevant pages at any position in the HITs and then randomly distributes pages brought in by the different pooling strategies.

**Measures** We look at *binary agreement* between the worker's label and the GS and at the agreement (Kendall's *tau*) on the comparative system ranking.

# 3. RESULTS AND ANALYSIS

**HIT designs** We see that the FullD labels agree significantly more with the GS labels than those from the SimpleD: 69% vs. 54% per HIT, and 67% vs. 51% per worker. This is further confirmed in Figure 2, showing that the FullD HITs attract workers who achieve significantly higher agreement levels with the GS labels. In Table 1 we summarize the correlations between the INEX ranking and the rankings based on the qrels from the crowdsourced labels. We observe a relatively high agreement between the FullD ranking and the INEX GS ranking across all metrics. The SimpleD and INEX rankings based on MAP and Bpref also correlate well ($\tau$ of 0.96 and 0.87), while the P@10 and nDCG@10 lead to poor correlations ($\tau$ of 0.34 and 0.02).

**Table 2: Agreement across different pooling strategies**

| Subset | rank-boost | top-n | answer-boost |
|---|---|---|---|
| FullD | 0.77 | 0.79 | 0.77 |
| SimpleD | 0.69 | 0.65 | 0.68 |

**Table 3: Impact of biased and random page order on system rank correlations with INEX ranking**

| Qrels | MAP | Bpref | P@10 | nDCG@10 |
|---|---|---|---|---|
| FullD-bias | 0.76 | 0.20 | 0.62 | 0.51 |
| FullD-rand | 0.78 | 0.63 | 0.93 | 0.81 |
| SimpleD-bias | 0.96 | 0.78 | 0.16 | -0.20 |
| SimpleD-rand | 0.94 | 0.82 | 0.44 | 0.24 |

**Pooling strategies** Table 2 shows no substantial difference between label accuracy levels for the three pooling strategies, which could be expected since workers are unaware of the origin of a page. Also the system-rankings over qrels based on the different pools (left out) are very similar.

**Page ordering** We see a significantly higher accuracy of labels for the random order of pages: FullD 73% vs. 65% and SimpleD 58% vs. 49%. This suggests that the ranking pattern of relevant pages influences the crowd workers' behavior, even when the known labels of these pages are not revealed. In Table 3 we show that the qrels obtained from HITs with random document ordering lead to higher correlation with the INEX ranking.

# 4. CONCLUSIONS

Our research investigates the use of crowdsourcing for collecting relevance judgments in IR tasks, such as book search, where the effort and cost of employing editorial staff is prohibitive. Crowdsourcing holds the promise of enabling large scale relevance assessments at modest costs, provided that care is taken when designing the HITs. We have three main findings. First, there is a need to measure the scope of impact by looking both at label accuracy and the resulting system ranking. In particular, the quality control element in the HIT design affects both label accuracy and system ranking, the pooling strategy only affects the system ranking, and ordering of pages in the HITs affects both label accuracy and system ranking. Second, we saw that the HIT design matters. Specifically, the quality control rich design is superior, and and implicit measures such as question dependencies can reduce the need for a gold set. Third, measuring the success of crowdsourcing is crucial. There is especially a need to use multiple metrics (MAP is too insensitive).

**Acknowledgment** This research was supported by the Netherlands Organization for Scientific Research (NWO projects # 612.066.513, 639.072.601, and 640.005.001) and by the European Community's Seventh Framework Program (FP7 2007/2013, Grant Agreement 270404).

# REFERENCES

[1] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. Crowdsourcing for book search evaluation: Impact of HIT design on comparative system ranking. In *SIGIR 2011*, pages 205–214, 2011. `http://doi.acm.org/10.1145/2009916.2009947`.