Crowdsourcing for Book Search Evaluation: Impact of HIT Design on Comparative System Ranking

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

[1] Microsoft Research, Cambridge UK, {v-gabkaz,natasamf}@microsoft.com
[2] University of Amsterdam, The Netherlands, {kamps,m.h.a.koolen}@uva.nl

ABSTRACT

The evaluation of information retrieval (IR) systems over special collections, such as large book repositories, is out of reach of traditional methods that rely upon editorial relevance judgments. Increasingly, the use of *crowdsourcing* to collect relevance labels has been regarded as a viable alternative that scales with modest costs. However, crowdsourcing suffers from undesirable worker practices and low quality contributions. 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.

Categories and Subject Descriptors: H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation (efficiency and effectiveness)*

General Terms: Experimentation, Measurement, Performance

Keywords: Prove it, Crowdsourcing Quality, Book Search.

1. INTRODUCTION

The evaluation and tuning of Information Retrieval (IR) systems based on the Cranfield paradigm [4, 27] requires purpose-built test collections, at the heart of which lie the human judgments that indicate the relevance of search results to a set of queries. With the ever increasing size and diversity of both the document collections and the query sets, gathering relevance labels by traditional methods, i.e., from a select group of trained experts, has become increasingly challenging [9]. This issue is especially prevalent in specialized search domains such as academic papers or books, which can support a range of tailored search tasks but also present additional complexities in IR evaluation. A good illustration is the INEX Book Track [13] which aims to provide a test bed for the evaluation of book search systems. The track reports on a range of issues related to the gathering of relevance labels [12, 13], one of which is the sheer effort of reviewing whole books and rendering relevance judgments for pages across a large number of retrieved books. While the INEX book collection comprises only 50,000 books, the effort to judge a single topic is estimated at 33 days if the assessor spent 95 minutes a day judging pages on that topic alone [12]. This estimate is based on a relatively shallow pool of 200 books per topic. The issue of scale in collecting human assessments is even more evident in the case of large online repositories that store millions of digitized books such as the Million Books repository[1] and the Google Books Library[2].

Recently, *crowdsourcing* [8] has emerged as a feasible approach to gathering relevance data in the context of IR evaluations [1–3, 7, 11, 15]. As such, it promises to offer a solution to the scalability problem that hinders traditional approaches based on editorial judgments, which has been the cornerstone of the IR evaluation since its conception at Cranfield [4] over 50 years ago. In general, crowdsourcing is a method of outsourcing work through an open call for contributions from members of a *crowd*, who are invited to carry out *Human Intelligence Tasks* (HITs) in exchange for micropayments, social recognition, or entertainment value. Crowdsourcing platforms, such as CrowdFlower[3] or Amazon's Mechanical Turk (AMT)[4] service, enable the gathering of vast amounts of data, such as relevance labels, from a large population of workers within a short period of time and at a relatively low cost.

However, the use of crowdsourcing presents a radical departure from the controlled conditions in which editorial judgments are collected. Thus, its use needs to be carefully examined before it can effectively supplement or replace the traditional methods. Indeed, the literature on crowdsourcing points to the poor quality of the resulting data [16, 22, 23, 30], which is often attributed to sloppy or even malicious workers and suboptimal HIT design [6, 11, 28]. Although best practices are gradually evolving, e.g., providing guidelines for the use of crowdsourcing in relevance as-

SIGIR'11, July 24–28, 2011, Beijing, China.

[1] http://www.ulib.org/
[2] http://books.google.com/
[3] http://crowdflower.com/
[4] http://www.mturk.com/

sessments [1, 3, 10, 14], the issues of attracting the 'right' workers and controlling their engagement in the crowdsourcing tasks remain a challenge.

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 use crowdsourcing to collect relevance judgments for book pages, imposing a considerable cognitive effort on the workers, and use the resulting labels to evaluate system performance on a *focused retrieval* task referred to as *Prove It* [13]. We study the elements of the crowdsourcing task design and tackle quality issues by incorporating a range of quality control methods. These include structured questionnaire flows using 'skip-logic' [18] and challenge-response tasks in the form of captchas adopted for digitized books [11]. We evaluate two different HIT designs, comparing how successful they are in motivating productive work behavior and examine the accuracy of the resulting relevance labels based on their agreement with the *gold standard* (GS) label set gathered from the participants of the INEX 2010 Book Track. Furthermore, we observe the relationship between the HIT designs, the label quality, and the resulting system rankings. The latter is motivated by the use of crowdsourcing as a means to an end, to scale up IR evaluation, the conclusions of which are based on the relative rankings of systems by given performance metrics. In our work we take the first step and observe how label accuracy impacts the IR system rankings, providing the basis for future work on quantifying the bias and uncertainty introduced by combining GS and crowd labels. We thus place the crowdsourcing of relevance labels within the larger context of IR evaluation.

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. More specifically, our experiments include two HIT designs, three pooling strategies, and two document ordering methods. Based on the analysis of the collected data, we provide insights on how design decisions influence both the raw label quality, i.e., agreement with GS, and the usefulness of crowdsourced relevance labels in IR evaluation, i.e., their impact on the system rankings. In particular, we study the implications of specific experimental conditions in order to answer the following research questions:

- What is the impact of quality control elements in the HIT design?
- What is the impact of the pooling strategy used to select pages to be judged?
- What is the impact of the ordering of pages in a HIT?
- What is the impact of consensus?
- What is the impact of selecting/rejecting workers based on agreement with the GS labels?

The rest of the paper is structured as follows. Section 2 discusses a use-case scenario for a large repository of scanned books: *Prove It*. We then review two possible strategies to evaluate the *Prove It* task: based on traditional editorial judgments and using crowdsourcing. We detail the design of a large-scale experiment in Section 3, aimed to study the effects of design decisions in crowdsourcing. Then, in Section 4, we analyze the impact of these mechanisms on the quality of the resulting relevance judgments and on the resulting system rankings for the official submissions to the INEX 2010 Book Track's *Prove It* task. Finally, in Section 5 we discuss our findings and draw conclusions.

2. BACKGROUND AND RELATED WORK

2.1 Prove It!

The INEX Book Track provides a forum for researchers to evaluate systems and methods for reading, searching, and navigating the contents of digitized books [13]. Among the IR scenarios supported by the track, since 2010, is the *Prove It* task which explores *focused retrieval* approaches in the context of book search. *Prove It* aims to investigate effective ways to retrieve relevant parts of books that can aid a user *in confirming or refuting a given factual claim*. This scenario is motivated by standard practices in academic authoring where authors use citations to refer to specific chapters or pages in other publications as evidence for an argument or a claim.

Reflecting the nature of the task, the topics of the INEX 2010 Book Track contain general knowledge factual claims. Using these topics, participating systems are required to retrieve and submit a ranked list of book pages per topic, estimated to provide information that can confirm or refute the topic claim or contain information that is related to the topic. Pages in the submitted runs are pooled and presented to assessors, the INEX participants, for relevance judgments using an online assessment system which enables users to browse, search, read and annotate books in the test collection [13]. Assessors are free to choose the topics they judge based on their expertise. For each topic, the system displays a list of books and, for each book, a list of pages compiled from the retrieval results of the participating IR systems. The books are ordered based on their rank positions and number of occurrences in the runs. As with the topics, assessors are free to choose which books to judge from the list. However, once inside a book, assessors are required to judge all listed pages. Assessors are also encouraged to discover further relevant pages, not included in the pools.

In this paper we us the *Prove It* task as the foundation for formulating the crowdsourcing tasks, implementing the crowdsourcing experiments, and analyzing the impact that aspects of the task design have on the evaluation of IR systems.

2.2 Evaluation Strategies

2.2.1 Traditional Editorial Judgments

The standard IR evaluation approach is based on the Cranfield paradigm [4] as adopted by the TREC [27] and other IR forums. This practice relies on test collections built under controlled conditions, consisting a set of documents, topics, and relevance judgments identifying which of the documents are relevant for a topic. The relevance judgments over the set of documents, selected by a given pooling method from the search results of individual systems, are made by (multiple) assessors following precise guidelines [27]. Studies of TREC systems evaluations have shown that relevance assessments are subjective in nature, leading to pairwise agreement levels of 70–80% on average between TREC assessors with high variability across topics [27, p.44]. The evaluation of systems under such variability of assessors and topics requires the use of a large set of topics and the comparison of system rankings under the exact same conditions on the same test collection. This is illustrated by several studies that replicated TREC relevance assessments with different judges [5] and found a significant disagreement on the relevance labels but considerable agreement on the resulting system rankings. Voorhees [26] also explored the impact of variations in relevance judgments on the system rankings. The use of pooling seems an important factor in the relative stability of system rankings, as even a set of randomly selected and labeled documents, assuming a good quality pool with relatively many relevant docu-

ments, can lead to a considerable correlation with a system ranking based on the editorial judgments [24].

Recently, there have been concerted efforts to find alternatives to explicit relevance judgments and infer relevance from the user's interaction with the system and the displayed content, e.g., from clicks on search results and hyperlinked pages. Although these approaches show reasonable levels of agreement with explicit judgments [21], it is still uncertain whether they can replace the traditional test collections with editorial relevance labels [9].

2.2.2 Crowdsourcing

The popularity and adoption of crowdsourcing have significantly increased with the emergence of widely accessible crowdsourcing platforms on the Internet. This has lowered the barrier for coordinating crowd workers to solve complex problems or engage in tasks that cannot be successfully completed without human intervention. Crowdsourcing has proven particularly useful for labeling large data sets such as acquiring image annotations and relevance assessments for search results [3, 7, 17, 29]. Among the successful examples to harness human resources on the Internet is the ESP game [25], which resulted in a large collection of labeled images by providing entertainment as the main value to its participants. An alternative approach to crowdsourcing is a low cost employment of workers as *mechanized labor* [20]. Crowdsourcing platforms, such as AMT, apply such a micro-payment job market model. Workers can select among a wide range of human intelligence tasks (HITs) that, usually, can be completed quickly for small financial rewards.

While popular, crowdsourcing is increasingly criticized for its poor output quality [16, 23, 30]. Indeed, quality assurance has surfaced as a major challenge and research topic in crowdsourcing [11, 15, 28]. Among main factors contributing to the suboptimal output are (1) workers' dishonest and careless behavior [6] and (2) poor task designs by the task requesters [7, 14]. Clearly, workers motivated by financial gains may aim to complete as many HITs as possible within a given time, which may reflect negatively on the quality of their work. Hence, it is important to devise methods for deterring dishonest and careless workers and detecting suboptimal output among completed tasks. By the same token, conscientious workers who are faced with an ambiguous task or inappropriate HIT design may produce work of inferior quality despite their best intentions. Thus, it is important to understand the principles of good HIT design.

Among quality control methods are tools of defensive task design, such as employing multi-level reviews of workers and leveraging reputation systems [20]. Other such tools include the use of trap questions [30], qualifying questions [6], gold standard data for which agreement can be measured [14, 15, 23], timing controls [10], and challenge-response tests (*captcha*) [11]. A common approach to correcting for unreliable workers is to build redundancy into the task design [20] and collect labels for the same item from different workers and aggregate them by applying a majority or consensus rule [22, 28]. Further steps involve analyzing the individual workers' output for consistency relative to both the label quality and the output of other workers. For example, Sheng et al. [22] model annotators' quality and show that repeated and selective labeling increases the overall quality of the output. Whitehill et al. [29] extend this work by considering the difficulty of the task and the ability of the annotators. Welinder et al. [28] combine both label and annotator quality in a generative Bayesian model.

All the above approaches use quality metrics that are directly focused on the crowdsourced data, such as the label accuracy relative to the GS labels. An exception is the work by Nowak and Rüger [17] who observe the effects of crowdsourced data on the comparative evaluation of systems for concept detection in images. Their study, on a set of 99 images, shows a high level of agreement between the system ranking based on "expert" labels and the ranking based on crowdsourced labels for 53 concepts.

In our research we take a dual approach and consider both the accuracy of the crowdsourced labels and their impact on the system performance rankings in an IR task with informational topics and relevance judgments, at the scale of an IR test collection. Moreover, by varying aspects of the HIT design, we observe the effects of labels from worker groups that are subject to different control mechanisms and thus exhibit different behaviors which may then be associated with different levels of trustworthiness.

3. APPROACH

In this section, we discuss our methodology for investigating the effectiveness of various aspects of HIT design as reflected in the resulting crowdsourced label accuracy and the ranking of IR systems. We devise experiments with different document pooling and ordering strategies and different sets of quality control mechanisms used to direct workers towards effective work practices and to deter dishonest workers. We hypothesize that different HIT designs lead to workers engagements at different levels of trustworthiness. Thus, varying the HIT designs enables us to observe how system rankings differ when we apply relevance labels from workers that are deemed more reliable based on higher accuracy of their labels and participation in more controlled HIT designs.

We use the problem of collecting relevance judgments for the INEX *Prove It* task as a motivation of the HITs for the workers.

3.1 Experiment Data

The data used in our experiments consists of 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[5] and 10 *Prove It* runs that were submitted to INEX 2010. Best Books run contain up to 100 books per topic, ranked in the order of estimated relevance. Each *Prove It* run contains a ranked list of up to 1,000 book pages per topic.

From the 83 *Prove It* topics in the 2010 test set, each containing a general knowledge factual claim, 29 topics were assessed by INEX participants. After removing topics with less than 10 known relevant pages, we arrive at the final set of 21 topics with an average of 169 judged pages per topic (total of 3,557 judged pages). This set comprises the GS data set in our experiments. The assigned relevance labels can take one of four values: 3 if the judged page contains information that *confirms* some aspect of the topic claim, 2 if the page contains information that *refutes* some aspect of the claim, 1 if the page contains information that relates to the claim but does not confirm nor refutes the claim explicitly, or 0 if the page is irrelevant. We use 'relevant' to mean labels in $\{1, 2, 3\}$.

3.2 Experiment Design

3.2.1 Pooling Strategy

In order to create the pools of book pages for relevance assessment in our HITs, we sample pages from the 10 *Prove It* runs submitted to INEX 2010. We experiment with three pooling strategies, including two approaches to boost the distribution of potentially relevant book pages in the assessment pool. Ensuring that more relevant documents are included has two benefits: it improves the cost effectiveness of obtaining complete relevance judgments for

[5]Ad hoc retrieval of whole books

the collection and improves the experience of the judges by reducing the number of non-relevant documents presented to them.

Top-n pool: We prepare an assessment pool of book pages following a standard top-n round-robin approach [19] to fill a fixed pool size of 100 pages from the *Prove It* runs, removing duplicates in the process.

Rank-boosted pool: We re-rank pages in the *Prove It* runs based on the book's highest rank and *popularity* across all the Best Books and the *Prove It* runs, i.e., the number of runs in which the book containing the page occurs. For *Prove It* runs we determine a book' rank based on the highest ranking page from that book. We order books with the same rank score based on their popularity score. Since in the *Prove It* task, multiple pages from the same book may be returned, they receive additional boost. This is the same method that was used at the INEX Book Track [13].

Answer-boosted pool: We re-rank pages in the *Prove It* runs based on their content similarity to the topic. We employ a simple similarity function based on coordination level or cohort matching to compute the scores and rank individual pages in this pool. We match the terms in the topic claim after removing stopwords and terms occurring in the query and the subject statement parts of the topic.

In the final step, we allocate 9 of the 10 pages per HIT by interleaving the three pooling strategies. This enables us to increase pool diversity and study the effects of pooling in our experiments. Our approach is similar to the interleaving strategies used by Radlinski et al. [21] in the Web context.

Although some overlap already exists between the resulting pools and the GS set due to the above sampling methods, we add to each HIT an additional page with known confirm/refute label (label 2 or 3). Thus, a single HIT of 10 pages includes at least 1 page with a known label, and up to 9 pages sampled from the three interleaved pooling methods. This results into 2,100 pages to be judged by crowd workers: 100 pages per topic with, roughly, 30 unique pages from each pooling method and additional pages with known labels purposely added to the HITs. Pooled pages included in a HIT are unique, but the known relevant page may be repeated across HITs. This leads to the total of 1,918 unique topic-book-page combinations: about 91 pages per 21 topics, covering 759 unique topic-book combinations, i.e., around 36 books per topic. Out of 10 pages per HIT there are, on average, 4.5 pages with a known label among which half, i.e., 2.2 are relevant pages with a known label.

3.2.2 Document Ordering

The study by Le et al. [15] shows that the distribution of relevant documents presented to workers during training affects their behavior later on: workers learn the distribution patterns and apply them in their judgments. Thus, we are interested in whether the ordering of book pages affects the resulting label accuracy and the system rankings. Unlike [15], we do not reveal the ground truth labels to the workers. Nevertheless, it stands to reason that having a relevant page at the top rank of each HIT may bias workers' relevance judgments, creating the expectation that the top document is always relevant. However, since the first page in a HIT may receive more attention by workers, using the first page to filter out dishonest or careless workers may be less effective. In order to study the impact of ordering, we consider two cases:

Biased order: We construct the HITs by preserving the order of pages produced by a given pooling approach, so based on decreasing expected relevance, and by inserting the known relevant page at the first position in the HIT.

Random order: We construct the HITs by first inserting the known relevant pages at any position in the HITs and then randomly distributing pages brought in by the different pooling strategies.

3.2.3 HIT Design and Quality Control

As discussed in Section 2.2.2, quality control is of paramount importance in crowdsourcing. In the context of the *Prove It* task, providing relevance judgments requires workers to (1) read and understand the claim that needs to be confirmed or refuted, (2) read the book page text, (3) identify relevant content, (4) make an inference, and (5) assign the appropriate relevance label from multiple possible labels. A worker's output is likely to be suboptimal if the worker fails to perform well any of the task stages. Thus, we need to devise control mechanisms that verify the proper worker engagement in each stage of the task in order to reduce careless behavior, including the extreme cases of dishonest workers' behavior. In our experiments, we build on established and emerging best practices in crowdsourcing [1, 6, 11, 14] and devise two types of HITs with different sets of quality control mechanisms.

Full design (FullD): This design, see Figure 1, controls all the stages of the task and explicitly pre-qualifies workers, restricting participation to those who completed over 100 HITs at a 95+% approval rate. Following experience that HIT titles influence workers' attention [7], in order to attract workers interested in a given topic, we group FullD HITs into 21 topic-specific batches, and include topic details in the title, description, and the keywords associated with the HITs. With the aim to deter sloppy workers, we explicitly warn workers that at least 60% of their labels need to agree with expert-provided labels in order to qualify for payment. As trap questions are useful for detecting careless workers who miss to read instructions [6], we include two simple trap questions: "Please tick here if you did NOT read the instructions" at the top of the form, and "I did not pay attention" as a relevance label option for each page. To test if workers have read and understood the factual claim for which they are to judge the book pages in the HIT, we ask qualifying questions for which a worker can either find the answer in the claim or respond based on their knowledge of the subject matter. To reduce the effectiveness of random clicking, we devised a structured series of questions based on skip-logic (Flow) to lead the worker through the assignment of relevance labels. The flow of question is such that answering the next question is dependent on the answer given to the previous question. Since only certain combinations of answers make sense, we can use this mechanism to detect careless or random responses. Captcha is commonly used to detect human input in online forms. We adopt it as a way to force workers to perform specific actions, e.g., read the content of a book page, and to verify that the action was indeed completed. For example, we ask workers to enter the first word of the sentence that confirms or refutes the claim. We include a captcha in each possible path of the Flow.

Simple design (SimpleD): This design includes a smaller number of quality control mechanisms and does not impose restrictions on the workers who can participate. Unlike the FullD version, all HITs are packaged into a single batch, using the same generic HIT title, description and keywords. The design includes only one trap question to control for random assignment of relevance labels. 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 from the FullD version, i.e., question D reduces to multiple choice and E is removed.

Common to both designs is that in all the HITs we include pages with known relevance labels. We have the same number of pages with GS labels in both FullD and SimpleD tasks.

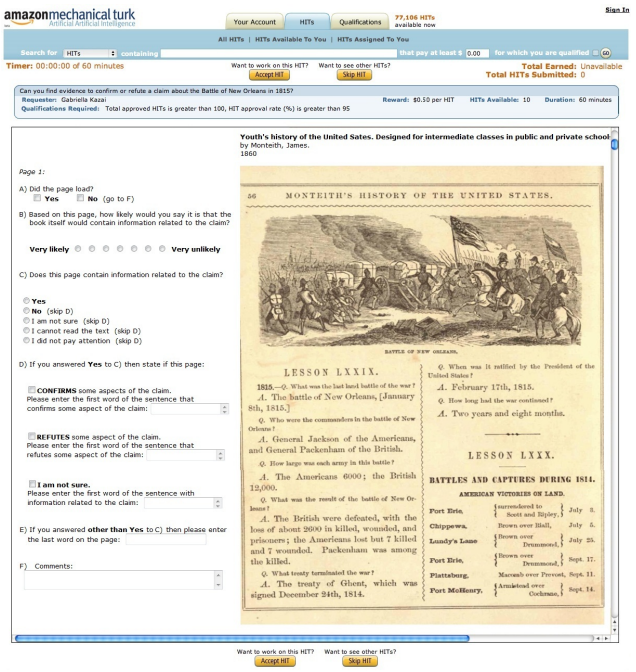


Figure 1: Part of a HIT showing question series to solicit relevance labels for book pages from workers on AMT: Full design.

While pay is an important influencer on workers' participation in AMT [16], in our experiments we aim to keep the pay per unit of effort constant across the designs. Based on our estimate that a FullD HIT requires twice the effort, in terms of time needed to complete, as a SimpleD HIT, we set payment at $.50 for FullD HITs and $.25 for SimpleD HITs, each comprising the relevance assessments of 10 pages.

3.2.4 Experiment Grid

Our goal is to investigate the effectiveness of crowdsourcing as we vary the pooling strategy and page ordering on one side and the quality assurance mechanisms that control the workers' engagement on the other. We create four batches of HIT experiments, see Table 1: 1) FullD-bias: Full design with biased ordering of pages; 2) FullD-rand: Full design with random ordering; 3) SimpleD-bias: Simple design with biased ordering; 4) SimpleD-rand: Simple design with random ordering. The interleaved pooling strategy is common across the experiments. We split the topic set between the biased and random batches for the FullD (10 and 11 topics, resp.), but run all 21 topics in the SimpleD task for both page orderings. For each FullD HIT we collect labels from three workers while for the SimpleD HITs we limit to one worker per ordering condition. Thus, we collect two and three labels per page for all 21 topics in SimpleD and FullD, respectively.

3.2.5 Measures

In order to assess the quality of the crowdsourced labels, CS, we rely upon the agreement of the labels with the GS assessments provided by the INEX 2010 Book Track, GS ={3,2,1,0}. We introduce two measures:

- *Exact Agreement (EA)*: Agreement on the exact degree of relevance, i.e., $CS = GS$,
- *Binary Agreement (BA)*: Either the page is non-relevant ($CS = 0$ and $GS = 0$) or relevant ($CS > 0$ and $GS > 0$) to the topic of the claim.

To investigate the impact on system rankings, we create test collections (qrels) from the sets of labels obtained from the different crowdsourcing experiments and evaluate the *Prove It* runs using standard metrics, e.g., mean average precision (MAP). We compare the resulting system rankings with the system ranking derived using the GS set as our test collection. In our experiment setup, a high correlation between system rankings based on the GS qrel and the qrels derived from the crowdsourced labels over the documents in the rank-boosted pool would mean that crowdsourcing leads to a comparable IR evaluation outcome as traditional methods.

4. ANALYSIS AND DISCUSSION

In this section we analyze the experimental results in two stages. We first observe the label accuracy against the gold set to understand how aspects of the HIT design can affect the accuracy of the crowdsourced labels. In the second stage we consider the comparative system rankings resulting from the crowdsourced relevance judgments, and look at the impact of the HIT design on the resulting test collections.

4.1 Impact on Label Accuracy

4.1.1 Impact of Quality Controls

Table 2 provides statistics on the number and the accuracy of relevance labels obtained under the different experiment conditions. We see that the FullD HITs yield considerably more labels per HIT and per worker than SimpleD (8.72 vs. 7.97 per HIT, and 45 vs. 40 per worker). In addition, the collected labels agree significantly more with the GS labels than those from the SimpleD HITs for both BA and EA measures (e.g., BA is 69% vs. 54% per HIT, and 67% vs. 51% per worker). This is further confirmed in Figure 2, showing the probability density distribution of workers across different agreement levels with the GS labels. We note a striking difference between the FullD and SimpleD HIT designs. The FullD HITs attract workers who achieve significantly higher agreement levels with the GS labels. We, thus, conclude that various aspects of the FullD design, from more informative HIT titles to flow questionnaires, induce more desirable behavior and more trustworthy workers' engagements. This is reflected in the significantly higher percentage of responses to the flow questions (Flow) for workers in FullD than in SimpleD (e.g., 79% vs. 67%), see Table 2.

4.1.2 Refining Relevance Labels

Consolidating multiple relevance judgments. Across HITs, we collected multiple labels per page, with the aim to mitigate possible noise from careless or dishonest workers by establishing internal agreement among workers through a majority vote. Thus, for every page with multiple relevance judgments, we determine the most popular label and, in case of a tie, keep the label with the lowest value among {3-confirms, 2- refutes, 1-relevant, and 0-irrelevant}. Other choices, such as keeping the label with the highest value or a random one, resulted in very similar outcomes.

As before, we use EA and BA to measure the accuracy of the resulting labels. It transpires that the FullD majority vote labels achieve 74% EA and 78% BA accuracy levels with the GS labels. This is significantly higher than the SimpleD majority vote labels with 61% EA and 68% BA levels. We note that these statistics are substantially higher than the mean agreement per HIT in Table 2: 62% EA and 69% BA for FullD and 44% EA and 54% BA for SimpleD.

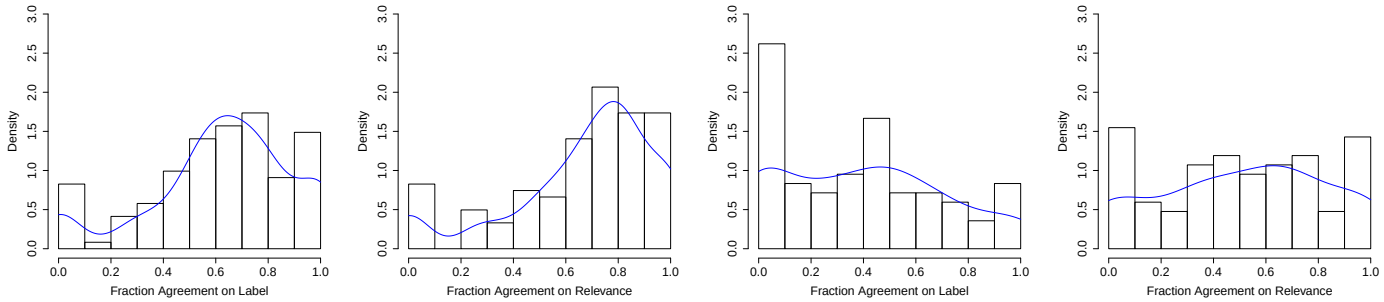
It is interesting to note that the accuracy of SimpleD labels is improved by the majority rule substantially more than for the FullD design. Furthermore, the accuracy of the consolidated labels of the

Table 1: Batches of HITs with different task parameter settings

Batch	HIT Design	Ordering	#Topics	#Workers per HIT	Total cost (incl. Amazon's 10% commission)
FullD-bias	Full design	Biased	10	3	(10 x 10 x 3 x $.50 x 1.10)=$165.00
FullD-rand	Full design	Random	11	3	(11 x 10 x 3 x $.50 x 1.10)=$181.50
SimpleD-bias	Simple design	Biased	21	1	(21 x 10 x 1 x $.25 x 1.10)=$57.75
SimpleD-rand	Simple design	Random	21	1	(21 x 10 x 1 x $.25 x 1.10)=$57.75

Table 2: Statistics per HIT (top) and Worker (bottom) over the batches and designs: number of completed HITs/number of unique workers; the total worker time in seconds; the total number of relevance labels; the exact (EA) and binary agreement (BA) over the GS labels; the fraction of pages respecting the questionnaire flow; and the fraction of pages with a captcha (FullD only). We test for significant differences between pairs of subsets using a two sample t-test (two-tailed, † for $p < 0.05$ and ‡ for $p < 0.01$)

Subset	# HITs or # Workers	Worker Time (s) mean	Worker Time (s) sd	#Labels mean	#Labels sd	%EA mean	%EA sd	%BA mean	%BA sd	%Flow mean	%Flow sd	%Captcha mean	%Captcha sd
FullD	630	876‡	645	8.72‡	2.46	0.62‡	0.32	0.69‡	0.32	0.85‡	0.25	0.26	0.20
SimpleD	420	395‡	432	7.97‡	2.97	0.44‡	0.34	0.54‡	0.33	0.75‡	0.31	–	–
FullD-bias	300	904	674	8.55	2.60	0.60	0.35	0.65‡	0.35	0.83	0.26	0.21‡	0.19
FullD-rand	330	851	618	8.88	2.32	0.65	0.29	0.73‡	0.28	0.86	0.23	0.30‡	0.21
SimpleD-bias	210	394	522	7.71	3.57	0.41	0.34	0.49‡	0.35	0.72	0.36	–	–
SimpleD-rand	210	396	317	8.23	2.19	0.48	0.33	0.58‡	0.31	0.77	0.25	–	–
All	1050	684	616	8.42	2.70	0.55	0.34	0.63	0.33	0.81	0.28	–	–
FullD	121	4562‡	7512	45.41	81.11	0.62‡	0.27	0.67‡	0.28	0.79‡	0.26	0.25	0.19
SimpleD	84	1976‡	2678	39.85	104.91	0.40‡	0.32	0.51‡	0.32	0.67‡	0.32	–	–
FullD-bias	70	3873	4539	36.68	50.38	0.62	0.27	0.66	0.27	0.81	0.23	0.20‡	0.15
FullD-rand	86	3265	4849	34.08	52.24	0.63	0.26	0.69	0.27	0.80	0.26	0.30‡	0.20
SimpleD-bias	42	1972	2849	38.59	130.75	0.36	0.35	0.45	0.33	0.59†	0.34	–	–
SimpleD-rand	45	1847	2298	38.38	69.29	0.45	0.29	0.58	0.30	0.75†	0.26	–	–
All	194	3700	7100	45.56	100.13	0.53	0.31	0.60	0.30	0.74	0.29	–	–



(a) Full Design (EA) (b) Full Design (BA) (c) Simple Design (EA) (d) Simple Design (BA)

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

SimpleD HITs almost reaches the average accuracy of the FullD HITs with a single label assignment. This suggests that there might be a cost trade-off that favors the use of simple HITs with multiple labels over the complex HITs with a single label collection. On the other hand, the effectiveness of majority rule is a function of not only the number of labelers but their individual labeling quality [22], which is higher in the FullD HITs. Overall, by collecting multiple relevance labels per page and applying majority voting, we arrive at a highly accurate set of crowdsourced labels.

Removing workers with low label accuracy. In Figure 3 we simulate the effect of rejecting workers below a given level of accuracy over the GS labels. We plot the agreement level for the resulting majority vote labels and show how filtering out workers with low accuracy labels increases the GS agreement for the remaining labels. Essentially, the agreement stays unchanged until the minimal accuracy of the workers reaches 40%. Note that substantially more workers are removed for SimpleD than FullD during this process, see Figure 2. Overall, we see that consensus is an effective way of eliminating potentially suboptimal contributions of low quality workers.

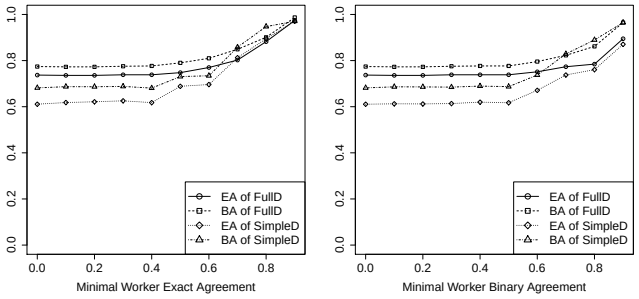


Figure 3: Agreement of majority vote label with the GS labels, after removing workers below n% exact agreement (left) or binary agreement (right).

Table 3: Exact (EA) and binary (BA) agreement across different pooling strategies

Subset	rank-boost		top-n-pages		answer-boost	
	EA	BA	EA	BA	EA	BA
FullD	0.73	0.77	0.74	0.79	0.70	0.77
SimpleD	0.62	0.69	0.56	0.65	0.58	0.68
All	0.75	0.79	0.73	0.80	0.74	0.85

Table 4: Unique pages and unique relevant pages assessed in HITs of different designs

Subset	Pages	rank-boost		top-n-pages		answer-boost	
		Uniq	Rel	Uniq	Rel	Uniq	Rel
FullD	1873	489	66	506	41	554	161
SimpleD	1816	473	168	489	131	531	202
All	1892	489	82	511	56	563	172

4.1.3 Impact of Pooling and Ordering Strategies

Comparing the impact of the biased and random order of pages in the FullD and SimpleD (shown in Table 2), we see a higher accuracy of labels for the random order of pages (significantly higher for BA over HITs using two sampled t-test with $p < 0.05$). 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 (as it was in the study by Le et al. [15]). There were two differences between the ordering strategies. The biased order HITs always start with a page that is relevant according to the GS, which is useful for priming workers. The biased order HITs were extracted in pooling order leading to a decreasing number of relevant pages per HITs. This variation leads to "unexpected" distributions of labels per HIT, ranging from all relevant to only 1 relevant page, which may influence workers' judgments rating less relevant pages as non-relevant, or workers looking for plausible patterns in the label distribution.

Since all the HITs were created by interleaving pages from three different pools, we can consider the accuracy of crowdsourced labels per individual pools. We restrict our attention to the pages that occur among the top 100 pages in the pool. Table 3 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. Table 4 shows the number of unique pages contributed by each pooling strategy and the number of relevant pages with labels obtained through majority voting. The number of unique pages from each pool is comparable. The answer-boost pooling leads to the highest number of unique and relevant pages, with a relatively high fraction of the unique pages judged relevant.

4.1.4 Factors Impacting Accuracy

The use of GS labels to measure the accuracy of the HITs output and infer the trustworthiness of the workers implies that a substantial effort is spent on the redundant assessment of pages with known relevance labels. Thus, it is of interest to identify other factors that are predictive of the label accuracy.

Table 5 shows the Pearson correlations between the label accuracy, in terms of the exact agreement (EA) with the GS labels, and several characteristics of the crowdsourcing engagement observed in the experiments. Correlations with the binary agreement (BA) are very similar and thus not shown.

We note that the total number of HITs completed by a worker provides no clues about the level of label accuracy ($r = 0.04$). Similarly the average time spent on a HIT is only weakly correlated with accuracy. However, the correlation between the EA and the number of labels produced by the workers ($r = 0.63$) is rather strong, indicating that dishonest and careless workers tend to skip parts of the HITs. As shown in Table 2, that happens significantly more often with the SimpleD HITs than with the FullD. The structure of the flow questionnaires (Flow) has a similarly high correlation with the EA accuracy ($r = 0.65$), implying that workers who respect dependencies among questions provide accurate labels. Interestingly, the captchas that were part of the FullD only, correlates less strongly with the EA. In fact, captchas were filled out only in 26% of the FullD HITs, see in Table 2. We speculate that their frequency and placement in the HIT design is suboptimal and require modifications. Overall, amongst the HIT factors, the completion of the flow questionnaires seems to be the best predictor of the produced label quality. That is expected since, by design, they prevent missing labels and ensure that appropriate logic is applied during label assignments. The latter is hard to mimic by random clicking and careless filling of the HIT forms.

Table 5: Pearson correlations between the label accuracy, in terms of exact agreement (EA) with GS labels, and HIT characteristics: the number of HITs per worker; the average time spent per HIT; the fraction of acquired relevance labels; the fraction of HITs with completed questionnaire flow; and the fraction of filled-in captchas (t-test, two-tailed, † for $p < 0.05$ and ‡ for $p < 0.01$).

Subset	#HITs	Time	Labels	Flow	Captcha
FullD	0.01	0.17	$0.72^{\ddagger}$	$0.71^{\ddagger}$	$0.29^{\ddagger}$
SimpleD	0.06	$0.26^{\dagger}$	$0.55^{\ddagger}$	$0.57^{\ddagger}$	–
FullD-bias	-0.06	0.04	$0.64^{\ddagger}$	$0.65^{\ddagger}$	$0.30^{\dagger}$
FullD-rand	0.03	0.20	$0.71^{\ddagger}$	$0.68^{\ddagger}$	$0.33^{\ddagger}$
SimpleD-bias	0.05	0.29	$0.51^{\ddagger}$	$0.51^{\ddagger}$	–
SimpleD-rand	0.05	$0.31^{\dagger}$	$0.55^{\ddagger}$	$0.62^{\ddagger}$	–
All	0.04	$0.30^{\ddagger}$	$0.63^{\ddagger}$	$0.65^{\ddagger}$	–

4.2 Impact on System Rankings

The above analysis shows that differences in the HIT design can lead to considerable variations in the label accuracy. FullD HITs which incorporate multiple control mechanisms to encourage productive worker engagements led to a high label accuracy over the GS set. This suggests that the output of the FullD HITs could be considered more trustworthy than the output of the SimpleD HITs.

We now explore how the use of crowdsourced labels affects the IR system ranking. To each set of labels we apply the majority rule and create a set of relevance judgments, *qrels*, to evaluate and rank the official submissions of the *Prove It* task, performed as part of the INEX 2010 Book Track. We use four performance metrics: MAP, Bpref, P@10, and nDCG@10. MAP and Bpref characterize the overall ranking and their comparison provides insights into the impact of un-judged pages. P@10 and nDCG@10 focus on the search performance in the top 10 retrieved pages. We apply nDCG measure by considering pages labeled 'relevant' as grade 1 and pages labeled 'refutes' and 'confirms' as grade 2. We refer to the system ranking based on the GS labels as the *INEX ranking* and measure its correlation with the rankings generated by other qrels using the Kendall's tau coefficient.

4.2.1 Quality Control

In Table 6 we summarize the correlations between the INEX ranking and the rankings based on the qrels from the crowdsourced labels. We note that document sets judged in HITs may be different from the GS documents. Thus, the observed ranking correlations are computed over different document samples in some instances.

Table 6: System rank correlation between the different designs

$Qrels_1$	$Qrels_2$	MAP	Bpref	P@10	nDCG@10
INEX	FullD	0.76	0.45	0.85	0.73
INEX	SimpleD	0.96	0.87	0.34	0.02
FullD	SimpleD	0.81	0.36	0.32	-0.16

Table 7: 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

We observe a relatively high agreement between the FullD ranking and the INEX ranking across all metrics. The SimpleD and INEX rankings based on MAP and Bpref also correlate well (τ of 0.96 and 0.87, respectively), while the P@10 and nDCG@10 lead to poor correlations (τ of 0.34 and 0.02, respectively). In fact, the correlation of INEX and the SimpleD rankings on MAP is remarkably high (τ of 0.96), higher than for the FullD (τ of 0.76). From the distribution of the 'relevant' label across the HITs, we find SimpleD workers to be more lenient, classifying 37% of the judged pages as relevant. On the other hand, the FullD workers regarded only 23% of them as relevant. Since all the HITs involve documents pooled from the results of the participating systems, this observation is in accord with the findings by Soboroff et al. [24]. They noted that by selecting a random set of documents from a high quality pool and treating them as "pseudo-relevant" leads to system ranking that is well correlated with the GS ranking based on MAP.

Comparing the effect of FullD and SimpleD qrels, we see that the system rankings based on P@10 and nDCG@10 disagree considerably and show low correlation for Bpref.

Overall, while the difference in MAP and Bpref indicate some discrepancy in the rankings, the P@10 and the nDCG@10 metrics more strongly differentiate the effects of the two HIT designs on the resulting system rankings, based on the qrels from the SimpleD HITs and the more trusted FullD HITs. The difference due to the lower reliability labels are primarily detected through metrics that focus on the top ranked pages.

4.2.2 Removing Workers with Low Accuracy

Earlier, we observed a considerable variation in label accuracy across workers. However, we also saw that removing workers with low accuracy had little effect on the overall accuracy of the majority vote labels. In Figure 4 we show the impact that incrementally filtering out low accuracy workers has on the correlation with the INEX ranking. For FullD we observe the convergence of the resulting system rankings towards the INEX ranking as workers with lower EA or BA are removed. In the case of SimpleD, the trend seems to be opposite. We attribute this to the weakening of the pooling effect as the number of relevant documents is reduced [24].

4.2.3 Impact of Pooling and Ordering Strategies

We found that random ordering of documents in the HITs yields higher levels of label accuracy compared to the biased ordering. In Table 7 we show the impact that this has on the system rankings. Overall, the qrels obtained from HITs with random document ordering lead to higher correlation with the INEX ranking.

Finally, we observe the influence that the three pooling strategies have on the system rankings. From Table 8 we see that the rank-boosted pool leads to very high correlations with the INEX ranking based on MAP for both the FullD and SimpleD qrels (τ of 0.94). For FullD this is also the case for the Bpref (0.90) and P@10 (0.91) metrics (and to a lesser extent for P@10 with τ of 0.87). Since the GS sample of documents is selected based on the same rank-boosted pooling strategy, these qrels have the highest overlap with the INEX qrels. Thus, the high correlation of system rankings suggests that the qrels comprising crowdsourced labels essentially lead to the same performance evaluation as the GS labels. We note that the more trusted relevance judgments from the FullD HITs lead to higher correlations with the INEX ranking across all the pooling strategies. From Table 4 we learnt that a substantial number of unique relevant pages are brought in by the top-n and answer-boost pools, explaining the increased divergence with the INEX ranking.

Table 8: Impact of pooling strategy on system rank correlations with INEX ranking

Qrels	MAP	Bpref	P@10	nDCG@10
FullD rank-boost	0.94	0.90	0.91	0.87
FullD top-n-pages	0.89	0.85	0.76	0.82
FullD answer-boost	0.72	0.72	0.59	0.51
SimpleD rank-boost	0.94	0.78	0.88	0.78
SimpleD top-n-pages	0.84	0.47	0.33	0.16
SimpleD answer-boost	0.81	0.64	0.44	0.47

Table 9: System rank correlations with INEX ranking over official submissions (top) and extended set (bottom)

Qrels	MAP	Bpref	P@10	nDCG@10
FullD	0.76	0.45	0.85	0.73
FullD rank-boost	0.94	0.90	0.91	0.87
GS+FullD	0.90	0.73	0.92	0.82
GS+FullD rank-boost	0.96	1.00	0.91	0.82
FullD	0.80	0.34	0.17	0.12
FullD rank-boost	0.90	0.82	0.72	0.71
GS+FullD	0.89	0.63	0.85	0.72
GS+FullD rank-boost	0.92	0.69	0.80	0.84

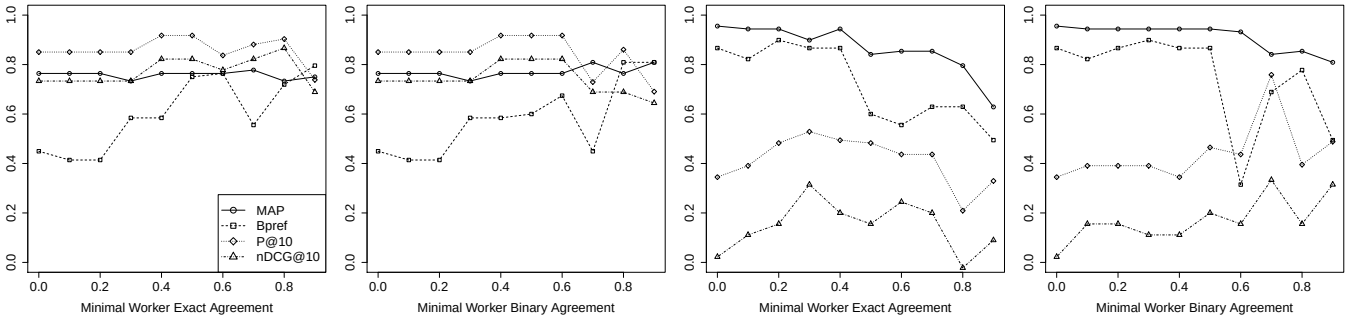
4.3 Evaluation of Prove It Systems

Crowdsourcing has a potential to improve evaluation of IR systems by scaling up relevance assessments and creating test collections with more complete judgments. However, as we employ crowd workers, we introduce uncertainty about the quality of the relevance labels. Our experiments were designed to investigate ways of using crowdsourcing to acquire reliable labels for a large portion of the test collection in order to approach the true system ranking based on complete relevance judgments by trusted judges.

In this section we investigate the use of the crowdsourced qrels to evaluate the *Prove It* runs of the INEX 2010 Book Track. We focus on the FullD HIT design where the tighter controls over the crowd engagement led to contributions from more trustworthy workers, with higher label accuracy. We conduct system evaluation (1) with qrels from FullD HITs only and (2) by merging FullD qrels with the GS qrels[6], expanding the GS with reliable labels. We also increase the set of 10 official submissions for the *Prove It* task with variants of the runs generated by re-ranking the retrieved pages based on the rank-boosting method.

Table 9 shows the correlations with the INEX ranking for the 10 official runs and the extended set of 20 runs. We evaluated systems

[6]We merge the qrels where we treat a page as relevant (or confirming/refuting a fact) whenever the INEX judge or the FullD majority vote says so.



(a) Full Design (EA) (b) Full Design (BA) (c) Simple Design (EA) (d) Simple Design (BA)

Figure 4: System rank correlation with INEX ranking after removing workers below n% exact or binary agreement.

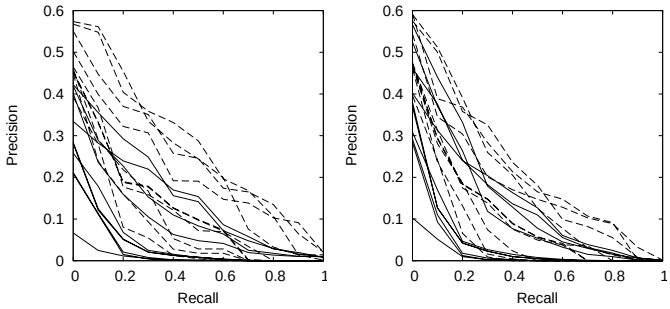


Figure 5: Interpolated precision over recall for the extended set of runs: gold set labels (left) and INEX+FullD labels (right).

using the FullD qrels and with its subset comprising pages from the rank-boosted pool only. As we saw before, the FullD qrels lead to slightly different system rankings from the INEX ranking since, by design, the crowdsourced document pool included pages outside the GS pool. As expected, the combined GS+FullD qrels, comprising the full GS label set leads to better correlation with the INEX ranking.

An interesting outlier is the low correlation of the extended system ranking based on the P@10 (τ of 0.17) and nDCG@10 (τ of 0.12) using the FullD qrels. These result from the additional pages in the FullD qrel that were contributed by the answer-boost and top-n pooling methods, which are ranked lower in the new runs that are based on the rank-boost re-ordering of documents.

Using the GS and the GS+FullD qrels we generate interpolated precision over recall plots in Figure 5, with the official (solid lines) and extended runs (dashed lines). We see that the extended runs dominate the top of the rankings evaluated against the GS labels. This is not surprising, considering that the new runs provide ranking of documents compatible with the pooling used to obtain the GS set. Evaluations using the FullD qrels only (not in Figure 5) and the combined set of GS+FullD qrels show that the performance of the official and the extended runs is closer to each other.

5. 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 highly skilled editorial staff is prohibitive. While pooling methods reduce the set of documents that need to be judged and make the problem more tractable, crowdsourcing holds the promise of enabling large scale relevance assessments at modest costs. However, the reliability of crowdsourced labels vary and one needs to design HITs to control the engagement of the crowd workers and the quality of their output.

In order to investigate how different aspects of the HIT design influence the label quality and the resulting system rankings, we conducted crowdsourcing experiments that involved: i) various quality control mechanisms embedded in the two HIT designs, FullD and SimpleD, ii) three interleaved pooling strategies for selecting pages to be assessed, and iii) two types of page ordering strategies, the biased and the random ordering within the HITs.

Our quality control approach was tailored to the *Prove It* relevance assessment task which involves several stages, from reading and understanding the topic claim, identifying the supporting or refuting evidence in the text, and then assigning a relevance label. Failing to complete any part of the task properly is likely to lead to suboptimal labels. Thus, we applied combinations of quality control mechanisms in order to encourage productive behavior across all the stages of the task. We characterized the HIT designs based on the label accuracy of the crowd workers over the pages with known relevance labels. This served as an indicator of the design's effectiveness in deterring careless and dishonest behavior, while attracting trustworthy workers who deliver reliable relevance judgments.

From our analysis of the collected labels we found that:

- The full design (FullD), with a rich set of quality control mechanisms, leads to significantly higher label quality in terms of agreement with the gold set labels.
- The random page ordering in the HITs leads to significantly higher label accuracy than the biased order where a known relevant page is likely to be placed at the top.
- Consensus over multiple judgments leads to more reliable labels, while filtering out workers with low accuracy leads only to a small increase in the label quality.
- The completion rate of the questionnaire flow and the fraction of obtained labels provide good indicators of the label quality. Flow is particularly effective in complex tasks since it enables both deterring and detecting suboptimal behavior, such as random clicking or filling of the HIT form.

From our analysis of the system rankings that resulted from the collected relevance judgments we conclude that:

- The choice of HIT design has a significant impact on the system rankings. FullD HITs, with a rich set of quality control mechanisms, lead to the highest correlation with the system ranking based on the gold standard set, the GS system ranking.
- System rankings based on MAP are highly correlated with the GS system ranking across the HIT designs. This suggests that MAP is not a good differentiator of the crowdsourced label quality. Measures like P@10 and nDCG@10 are more susceptible to the varying accuracy of the labels and the resulting qrels. This points to the use of multiple metrics

when evaluating the effectiveness of crowdsourcing through system rankings.

- Filtering out workers with low label accuracy reduces the pooling effect ([24]) that we observed when comparing system rankings based on MAP. By removing workers with low label accuracy, we observe a slight increase in correlation with the GS system ranking based on MAP for the FullD HITs and a decrease in correlation for the SimpleD HITs. The latter is the result of the diminished pooling effect since the number of 'relevant' labels is reduced and, with that, the possibility of achieving high correlation that can be otherwise be attained for any random sample of high-quality pooled documents.

Overall, our research suggests that crowdsourcing provides a useful means of acquiring relevance judgments for the evaluation of IR systems, assuming that reasonable care is taken when designing the HITs and the methods for refining relevance labels, e.g., based on consensus across multiple labels. From the experiments that compare results of crowd workers and editorial judges over the same ranking of pooled documents (rank-boosted), we found that the correlation of resulting system rankings is high for the quality control rich HIT designs across performance metrics. This gives us confidence that the quality of test collections obtained from crowdsourcing can be sufficiently good to enable reliable system evaluation. However, our analysis also highlights the varying quality of the workers' engagement and the resulting labels and a danger of relying upon a single metric in system evaluation, such as MAP, to draw conclusions regarding the success of a crowdsourcing experiment. Thus, crowdsourcing remains an instrument that should be used with due care and that requires further investigation.

Acknowledgments

Jaap Kamps and Marijn Koolen were supported by the Netherlands Organization for Scientific Research (NWO, grant # 639.072.601).

REFERENCES

[1] O. Alonso and R. A. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *Advances in Information Retrieval – 33rd European Conference on IR Research (ECIR 2011)*, volume 6611 of *LNCS*, pages 153–164. Springer, 2011.

[2] O. Alonso and S. Mizzaro. Can we get rid of TREC assessors? using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 15–16, 2009.

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

[4] C. W. Cleverdon. The Cranfield tests on index language devices. *Aslib*, 19:173–192, 1967.

[5] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21st annual international ACM SIGIR conference*, SIGIR '98, pages 282–289, 1998.

[6] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor. Are your participants gaming the system?: screening mechanical turk workers. In *CHI*, pages 2399–2402. ACM, 2010.

[7] C. Grady and M. Lease. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 172–179, 2010.

[8] J. Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Publishing Group, 2008.

[9] J. Kamps, S. Geva, C. Peters, T. Sakai, A. Trotman, and E. Voorhees. Report on the SIGIR 2009 workshop on the future of IR evaluation. *SIGIR Forum*, 43(2):17–36, December 2009.

[10] A. Kapelner and D. Chandler. Preventing satisficing in online surveys: A 'kapcha' to ensure higher quality data. In *The World's First Conference on the Future of Distributed Work (CrowdConf2010)*, 2010.

[11] G. Kazai. In search of quality in crowdsourcing for search engine evaluation. In *Advances in Information Retrieval – 33rd European Conference on IR Research (ECIR 2011)*, volume 6611 of *LNCS*, pages 165–176. Springer, 2011.

[12] G. Kazai, N. Milic-Frayling, and J. Costello. Towards methods for the collective gathering and quality control of relevance assessments. In *Proceedings of the 32nd international ACM SIGIR conference*, SIGIR '09, pages 452–459, 2009. ACM.

[13] G. Kazai, M. Koolen, J. Kamps, A. Doucet, and M. Landoni. Overview of the INEX 2010 book track: Scaling up the evaluation using crowdsourcing. In *Comparative Evaluation of Focused Retrieval : 9th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2010)*, LNCS. Springer, 2011.

[14] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 453–456, 2008. ACM.

[15] J. Le, A. Edmonds, V. Hester, and L. Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, pages 21–26, 2010.

[16] W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09, pages 77–85, 2009. ACM.

[17] S. Nowak and S. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 557–566, 2010. ACM.

[18] A. N. Oppenheim. *Questionnaire Design, Interviewing and Attitude Measurement*. Pinter Publishers, London, 1992.

[19] B. Piwowarski and M. Lalmas. Providing consistent and exhaustive relevance assessments for XML retrieval evaluation. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM'04, pages 361–370, 2004. ACM.

[20] A. J. Quinn and B. B. Bederson. Human computation: A survey and taxonomy of a growing field. In *Proceedings of CHI 2011*, 2011.

[21] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *CIKM*, pages 43–52. ACM, 2008.

[22] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceeding of the 14th ACM SIGKDD international conference*, KDD '08, pages 614–622, 2008. ACM.

[23] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, 2008.

[24] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th annual international ACM SIGIR conference*, SIGIR '01, pages 66–73, 2001.

[25] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 319–326, 2004. ACM.

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

[27] E. M. Voorhees and D. K. Harman, editors. *TREC: Experimentation and Evaluation in Information Retrieval*. MIT Press, 2005.

[28] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23*, pages 2424–2432, 2010.

[29] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of the 2009 Neural Information Processing Systems (NIPS) Conference*, 2009.

[30] D. Zhu and B. Carterette. An analysis of assessor behavior in crowdsourced preference judgments. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, pages 17–20, 2010.