A Little Competition Never Hurt Anyone's Relevance Assessments

Yuan Jin, Mark J. Carman
Faculty of Information Technology
Monash University
{yuan.jin,mark.carman}@monash.edu

Lexing Xie
Research School of Computer Science
Australian National University
lexing.xie@anu.edu.au

Abstract

This paper investigates the effect of real-time performance feedback and competition on the accuracy of crowd-workers for a document relevance assessment task. Through a series of controlled tests, we show that displaying a leaderboard to crowd-workers can motivate them to improve their performance, providing a bonus is offered to the best performing workers. This effect is observed even when test questions are used to enforce quality control during task completion.

1 Introduction

Crowd-sourcing involves enlisting the skills and expertise of many individuals (usually over the Internet) to achieve tasks that require human-level intelligence. Paid crowd-sourcing platforms, such as Amazon's Mechanical Turk[1] and CrowdFlower[2], allow for the low-cost outsourcing of labelling tasks such as judging document relevance to Web search queries.

In this paper we investigate paid crowd-sourcing applied for the collection of relevance judgements for Information Retrieval tasks and look to better understand the role performance feedback and competition can play in motivating crowd-workers to provide high-quality judgements. More specifically, we seek to systematically answer the following *research questions*:

- Does providing *real-time feedback* to crowd-workers regarding their relevance judging accuracy affect their average performance? And if so, does their performance improve or worsen?
- Does the competition between crowd-workers that results from *providing a leaderboard* to them affect their performance? And if so, does it make the performance of workers better, worse or both (i.e. more varied)?
- Do crowd-workers need to be incentivised by financial rewards (i.e. bonuses/prizes to the best performers) to improve their performance?
- If crowd-workers are already motivated to perform well through competition, are test questions (i.e. the vetting of workers using pre-task quizes and in-task assessments) still necessary in order to motivate them to provide quality judgements?

Given these research questions, the main *contributions* of the paper can be summarised as follows:

- We show that it is straightforward to augment a standard crowdsourcing platform with basic real-time feedback and leaderboard functionality while preserving privacy and anonymity of the crowd-workers.
- We introduce an experimental design based on standard A/B/n testing [KLSH09] that assigns crowd-workers directly to control and treatment groups. By further partitioning documents across the experiments we look to prevent contamination both within and between experiments.
- We demonstrate via controlled experiments that providing real-time feedback to crowd-workers via a leaderboard while offering them a financial reward for high achievement likely motivates them to provide better relevance judgements than they would otherwise, (i.e. if no leaderboard were present).

In: F. Hopfgartner, G. Kazai, U. Kruschwitz, and M. Meder (eds.): Proceedings of the GamifIR 2016 Workshop, Pisa, Italy, 21-July-2016, published at http://ceur-ws.org

[1] https://www.mturk.com/mturk/welcome

[2] https://www.crowdflower.com/

- We demonstrate that even when control/test questions are used to vet and remove low performing crowd-workers, the positive effects of real-time feedback and competition are still present.

The paper is structured as follows. We first discuss related work in Section 2, followed by describing the experimental setup and three controlled experiments in Section 3. We draw the final conclusions in Section 4.

2 Related Work

Alonso and Mizzero [AM09] demonstrated in 2009 that crowd-workers on a popular crowd-sourcing platform can be effective for collecting relevance judgements for Information Retrieval evaluations with the workers in some cases being as precise as TREC relevance assessors. Since then there has been a great deal of work in the Information Retrieval community investigating the efficacy of crowd-sourcing for collecting relevance judgements, with guides being developed [ABY11] and three years of TREC competitions organised [SKL12].

Grady and Lease [GL10] investigated the importance of user-interface features and their effect on the quality of relevance judgements collected. In particular, they investigated human factors such as whether offering to pay a bonus for well justified responses (they asked users to explain their judgements) resulted in higher accuracy judgements. They found that accuracy did indeed markedly increase as a result of the offer. (We find similar effects and control for them in our experiments.)

A recent survey of gamification ideas (such as providing points and leaderboards to users) applied to Information Retrieval tasks was provided by Muntean and Nardini [MN15]. Of particular note is the work by Eickhoff et al. [EHdVS12], who developed a game-based interface for collecting relevance judgements in an attempt to improve the quality and/or reduce the cost of crowdsourced relevance judgements. Despite similar goals to ours, their game-based interface was not integrated within a standard crowd-sourcing environment, and relied on quite different mechanisms for establishing which documents were relevant to which queries, thus making the work not directly comparable to ours.

Harris [Har14] also investigated gamification ideas with respect to relevance judgement and in particular investigated the difference between user's own perception of relevance and their perception of the "average user's" opinion of what constitutes relevance. Harris made use of a leaderboard at the start and end of the game, but did not provide real-time feedback about performance via the leaderboard directly to the individual crowd-workers like our work does.

In addition to developing practical expertise in the collection of crowdsourced relevance judgements, researchers have developed specialised algorithms for identifying the most likely correct judgement (or a posterior over relevance judgements) for a query-document pair based on a conflicting set of judgements from multiple users (see for example [MPA15]). Other work has looked at making direct use of crowd-sourced relevance judgements in order to train rank-learning systems (see for example [MWA+13]).

Despite all of this previous research, we are unaware of any work that has tested specifically the conjecture that providing real-time feedback to crowd-workers regarding their relevance judging performance can improve the same. Moreover, the use of leaderboards to motivate crowd-workers to provide better relevance judgements has not been previously investigated to the best of our knowledge.

3 Experiments

In this section, we detail the three experiments performed to address the research questions specified in Section 1.

3.1 Dataset

We made use of the "assigned-judged.balanced" data subset[3] from the TREC 2011 crowd-sourcing track[4] with 750 expert judgements made by NIST assessors for 60 queries and the ClueWeb09 collection[5] from which the documents judged for relevance were drawn.

3.2 Experiment Design

In each experiment, we collected relevance judgements for a crowd-sourcing task undertaken by workers on the Crowdflower platform. Within each task, we performed A/B/n testing of some control and treatment groups. We followed standard techniques for online experimentation [KLSH09], dividing workers as equally as possible across different control and treatment groups in each task. Moreover, we randomly divided the chosen data subset into three disjoint sets of document-query pairs in order to prevent crowd-workers who might partake in multiple experiments from ever judging the same document-query pair twice. Within an experiment, the same set of query-document pairs were shown to all crowd-workers across (and within) the control and treatment groups in order to remove variability due to (query-document pair) judgement difficulty.

[3] This data subset is balanced across three relevance levels (i.e. highly relevant, relevant and non-relevant)

[4] https://sites.google.com/site/treccrowd/2011

[5] http://lemurproject.org/clueweb09/

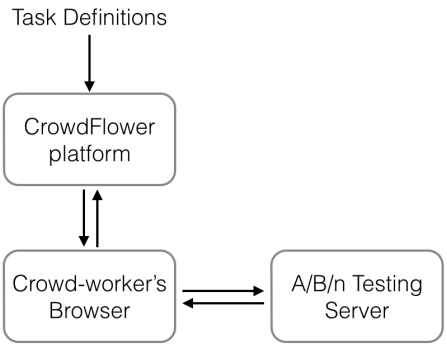

Figure 1: Architecture of experiment environment. Crowd-workers are randomly assigned to a particular group after joining the task on CrowdFlower, and see the same version of the interface for each subsequent interaction from their browsers. The server updates their performance statistics (or rank positions) once each page of judgements was completed.

The Crowdflower platform is designed for ease of use in creating the crowd-sourcing tasks, but not for the A/B/n testing within each task. Thus in order to perform controlled experiments we employed the experiment architecture shown in Figure 1. The crowd-workers first join our tasks held on CrowdFlower platform, causing a task page to load on their browser. At this point, client-side Javascript sends a request for additional page content to the A/B/n testing server. The server then randomly allocates each individual to one of the control and treatment groups for the duration of the experiment[6], and returns content to be inserted into the task page.

The inserted content provided different information to crowd-workers in different experimental groups. For example, the task page shown in Figure 2 contains the performance statistic (e.g. labelling accuracy) for the worker and his/her own view on the shared leaderboard, while the task page shown in Figure 3 (for a different group) contains no inserted content. The performance statistics for each crowd-worker is updated by the server based on the ground-truth relevance judgements (from NIST assessors) at the completion of each task page (consisting of 5 relevance judgements). The crowd-worker's view on the leaderboard is updated each time they load *or refresh* the task page. The leaderboard shown for the various experiments contained: (i) the rank of the crowd-worker, (ii) anonymous usernames for each of the crowd-workers, (iii) the score for each worker (computed using metrics to be specified in the following sections), and (iv) the change

[6]We used random assignment rather than sequential assignment to prevent biases due to the order in which individuals join the tasks.

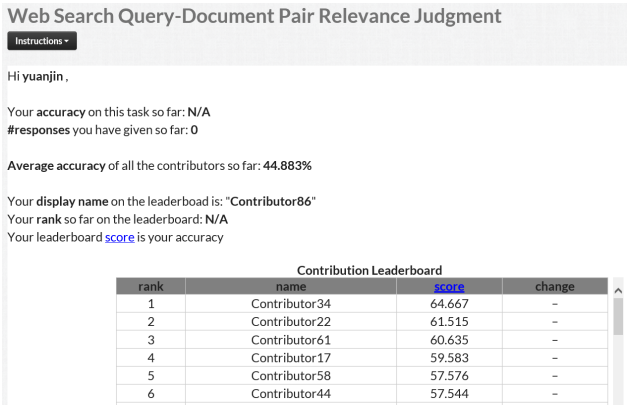

Figure 2: A screenshot of the interface shown to crowd-workers assigned to Treatment group 2 for Experiment 1, containing a leaderboard ranking contributors based on their labelling accuracy in percentage.

Figure 3: A screenshot of the interface shown to crowd-workers assigned to the control group for Experiment 1, which contained no additional information about the workers' performance.

in the rank of each worker since the previous refresh.

Crowd-workers were free to quit a task at any point, sometimes even without making any judgement, meaning that the number of workers can differ across different groups due to both random assignment and self-removal.

We adopted the setup used in the TREC 2011 crowd-sourcing track where a worker judged the relevance of 5 documents per query. The 5 documents per query were embedded as images in one page in random order for all the tasks except the one which will be specified in Section 3.5.

3.3 Experiment 1: Real-time Feedback

In the first experiment, we investigated whether providing real-time feedback to crowd-workers affected their performance. To do this we randomly assigned

crowd-workers to one of four groups:

- **Control Group:** *No performance feedback* was provided to the workers.
- **Treatment 1:** The workers were told their *current estimated accuracy* and the *current average estimated accuracy of all the workers* in the same group. (The latter information was provided for anchoring purposes, so that the workers knew whether their accuracy scores were relatively high or low.)
- **Treatment 2:** A *leaderboard* was provided to the workers showing their current ranking in the group scored by their estimated accuracy. On the board, workers were referred to as "Contributor$i$" where i denoted the order in which they were assigned to this group. Thus no personally identifying information (such as CrowdFlower usernames) was leaked between contributors. Moreover, workers got to see the changes in their ranking every time they reloaded current pages or loaded new ones.
- **Treatment 3:** Modify the scores for the workers on the leaderboard to be the product of their estimated accuracy and the numbers of responses they have made. The idea was to encourage workers to keep annotating in order to rank higher on the board.

For this experiment, each task consisted of labelling 5 documents for the same query as {*highly relevant*, *relevant*, *non-relevant*}. Crowd-workers could complete a maximum of 20 tasks each (i.e. label 100 documents) and each task was made available to 40 crowd-workers. The documents used to build the tasks were randomly sampled from the balanced data subset of the TREC2011 dataset.

In all experiments, we used the term "estimated accuracy", even though we calculated the accuracy based on the ground-truth relevance judgements, in order to be consistent with the general case where the ground-truth judgements are unknown and need to be estimated. Moreover, we applied Dirichlet smoothing to the accuracy estimate[7] $\hat{\mu}_i$:

$$\hat{\mu}_i = \frac{(\sum_{j=1}^{n_i} 1(r_{ij} = y_j)) + \alpha \frac{1}{3}}{n_i + \alpha} \quad (1)$$

Where n_i is the number responses from Worker i, r_{ij} denotes the individual's response to question j, y_j is the ground-truth response, and α is a smoothing parameter set to 5 so that the number of pseudo-counts equals the number of questions per page. The main reason for smoothing was to prevent crowd-workers from providing a small number of judgements on which they performed unusually well (e.g. judge all 5 documents correctly on the first page) and then quitting the task to win the competition.

Some contributors might have provided fewer relevance judgements than others because they (i) arrived late to the task when few pages were left or (ii) gave up early on. To prevent problems due to poor estimation of worker accuracy and/or weaker effects due to receiving feedback for shorter periods of time, data from workers who had provided less than 50 judgements were discarded across all groups (and were not included in the following analysis).

Table 1 shows experimental results across the four groups measured in terms of both micro and macro-averaged Accuracy[8]. The former is averaged equally

Table 1: Results across the four groups for Experiment 1. Accuracy is given as both Micro and Macro averages, with the latter being aggregated at the worker level. Crowd-workers who judged less than 50 documents were excluded from the analysis.

Group	# Workers	# Judgements	Accuracy (Micro)	Accuracy (Macro)
Control	44	3874	45.04%	45.39%
Treatment 1	40	3449	45.26%	43.20%
Treatment 2	48	3947	47.76%	47.86%
Treatment 3	39	3613	46.39%	47.06%

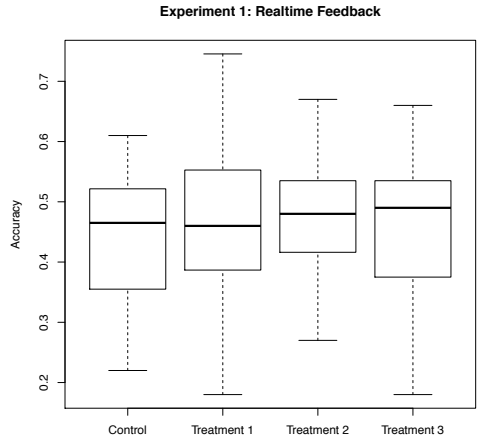

Figure 4: A boxplot showing the distribution of Accuracy values across the crowd-workers for different groups in Experiment 1.

[7] More formally, we employed an estimate of the posterior mean assuming a Binomial likelihood (over correct/incorrect answers) and a Beta prior with concentration parameter α and prior mean (probability of a correct answer) of 1/3 due to the balanced classes used.

[8] Accuracy was calculated treating the graded relevance levels {*highly relevant*, *relevant*, *non-relevant*}, as separate classes. Thus the assignment of label *highly relevant* to a document with

across all judgements and the latter equally across all crowd-workers. Due to the differing levels of ability/accuracy of the crowd-workers, the macro-average is the main measure of interest for determining changes in worker accuracy across the groups.

A boxplot in Figure 4 shows the distribution (median and inter-quartile ranges) of performance for workers in the various groups. We see that the spread of values is similar across the four groups.

To determine whether the differences in the mean performance across the groups were significant at the worker level, we employed a pairwise t-test (with non-pooled standard deviations and a Bonferroni correction for multiple comparisons). Given the test procedures, we did not find any significant difference between the groups[9], indicating that the effect size for the two treatments is small at best, i.e. there is little (if any) effect on performance that results from simply providing a leaderboard to crowd-workers and experiments with larger numbers of workers would be required to determine the size of the effect.

We also note from Table 1 that the score (the estimated number of correct judgements) used to rank crowd-workers in treatment 3 didn't appear to work as well as the score (the estimated accuracy) used in treatment 2 in terms of both micro and macro-averaged Accuracy. We conjecture that the reason for this might have been that new workers started very low on the leaderboard in the former case, and it took them a long time to rise up amongst the leaders, which caused apathy towards higher rankings.

3.4 Experiment 2: Adding a Bonus

Given the results of Experiment 1, we believed the crowd-workers needed to be further motivated and therefore moved to investigate whether incentivising the workers by offering to pay them a bonus at the end of the task affected their performance and if so, whether it caused them to perform better or worse.

More specifically, a $1 bonus was rewarded to those who ended up within the top 10 of the leaderboard. We followed the same setup used in Experiment 1 except the number of responses collected for each document-query pair for all the groups was now 30. The control and the treatment groups involved in this experiment are listed as follows:

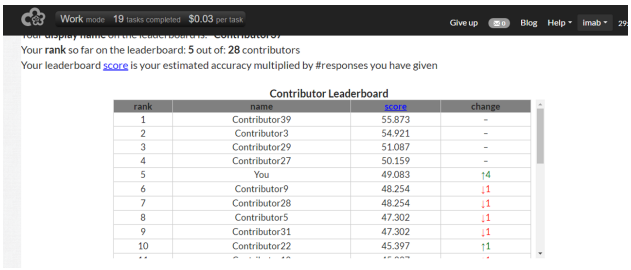

Figure 5: A screenshot of the interface shown to crowd-workers assigned to Treatment Group 2 in Experiment 2.

- **Control Group 1:** *Neither performance information nor bonus* was given to the workers assigned to this group.
- **Control Group 2:** *No performance information* was provided to the workers while they were performing their annotation, but they were informed that the *top 10* workers would be paid a $1 bonus at the completion of the task (once all the workers submitted their responses). Workers were also informed that their performance would be judged according to their *estimated numbers of correct responses*[10].
- **Treatment 1:** the configuration was the same as for Control group 2 except that a *leaderboard* was provided to the workers in the group showing their rankings in terms of the *Dirichlet-smoothed accuracy estimate*.
- **Treatment 2:** the configuration was the same as for Treatment group 1 except that the scores shown on the *leaderboard* were the *estimated number of correct responses* (as defined in Control Group 2) of the workers.

The results for the second experiment are shown in Table 2 with distributions over worker accuracies shown in Figure 6. Of note in the figure is the smaller interquartile range for Treatment group 2 with respect to the other groups (or the previous experiment). This may indicate that crowd-workers tend to perform more consistently when competing for bonuses on a leaderboard.

Treatment group 2 exhibits approximately 3% higher accuracy than the other groups for this experiment. However, pairwise T-Test results show no significant difference between the various groups at the 0.05 level after a Bonferroni correction[11]. The small-

true label *relevant* is considered incorrect. We repeated the analysis using Accuracy computed over binary relevance judgements, i.e. where the labels *highly relevant* and *relevant* were collapsed into the same class, finding similar results.

[9]The smallest P-value observed (before the Bonferroni correction) was 0.19 between Treatment 2 and the Control group, so not significant at the 0.05 confidence level even before correction for multiple comparisons.

[10]Estimated number of correct responses from Worker i is the product of the individual's Dirichlet-smoothed accuracy estimate and the number of responses he/she has made.

[11]Were it not for the correction for multiple comparisons, significant differences would have been claimed. Moreover, one-way

Table 2: Results for Experiment 2, where a bonus was offered to participants in Control group 2 and Treatment groups 1 and 2.

Group	# Workers	# Judgements	Accuracy (Micro)	(Macro)
Control 1	26	2419	41.71%	42.19%
Control 2	28	2624	42.11%	42.59%
Treatment 1	27	2665	43.83%	43.84%
Treatment 2	33	3139	46.54%	46.85%

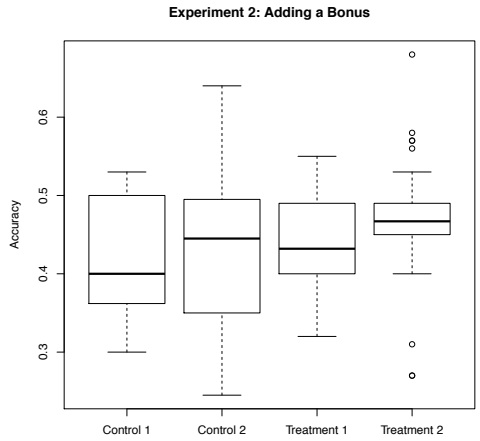

Figure 6: A boxplot of Accuracy across the different groups in Experiment 2 where a $1 bonus was paid to the top 10 contributors in Control group 2 and Treatment groups 1 and 2 at the end of the task.

est P-value is 0.14 between treatment 2 and control 1. Values indicate that significant differences would likely be observed for an experiment with a larger number of crowd-workers.

3.5 Experiment 3: Control Questions

Control/test questions are often used in crowdsourcing systems to check whether crowd-workers are (i) qualified for a certain task and (ii) constantly motivated to perform at a reasonable level throughout the task (i.e. not assigning random answers or same answers to all the questions[12]).

In CrowdFlower, conditions (i) and (ii) are implemented as respectively a quiz that workers must pass in order to embark on a task and task pages each of which contains one test question (amongst the other non-test ones).

In Experiment 3, we investigated what effect introducing such control/test questions would have on crowd-workers' performance of judging relevance on

ANOVA rejects the null that the group means are equal with a P-value of 0.019.

[12]This behaviour was observed in Experiment 1 for two workers in the Control group.

Table 3: Results for Experiment 3, where test questions were used (in both the control and the treatment groups) to vet and remove crowd-workers based on their corresponding accuracy.

Group	# Workers	# Judgements	Accuracy Micro	Macro
Control	40	3495	52.85%	52.53%
Treatment	45	4070	54.84%	54.82%

CrowdFlower and whether the real-time feedback and the competition functionality provided by the leaderboard were still useful for motivating the workers to annotate more and better given the certain level of quality control provided by the test questions.

More specifically, for each group in Experiment 3, we tried to collect 50 responses for each of the 100 task questions differing from those used in Experiments 1 and 2. The task was organised to comprise one quiz page which contained 5 test questions (not included in the task questions) and 20 task pages each of which contained one test question (included in the task questions) randomly inserted by CrowdFlower and four non-test ones randomised by us to appear on the same page. As a result, the 5 documents per page in Experiment 3 corresponded to different queries. There were in total 25 test questions (5 for the quiz and 20 for the test pages) and 80 non-test questions in this experiment. Thus a qualified crowd-worker could judge a maximum of 105 documents (5 in the quiz and 100 in the task itself). Moreover, we set the minimum accuracy of the test questions to 0.6, higher than the average Macro Accuracy (around 0.45) achieved by the groups in the previous experiments to allow the test questions to affect crowd-workers' annotation processes, while not so high that the workers get expelled from the task early on.

Only two groups were investigated in this experiment:

- **Control Group:** *Test questions* were used to select crowd-workers based on their corresponding accuracy, which in turn was provided back to the workers by CrowdFlower. No other performance information was provided, but workers were informed that the *top 10* would be paid a $1 bonus at the completion of the task.
- **Treatment 1:** The configuration was the same as for the Control group except that a *leaderboard* was provided to the workers showing their rankings in the group in terms of the *estimated number of correct responses* across all the documents judged (i.e. both test and non-test questions).

The results of the third experiment are shown in Table 3. We note that the overall accuracy across

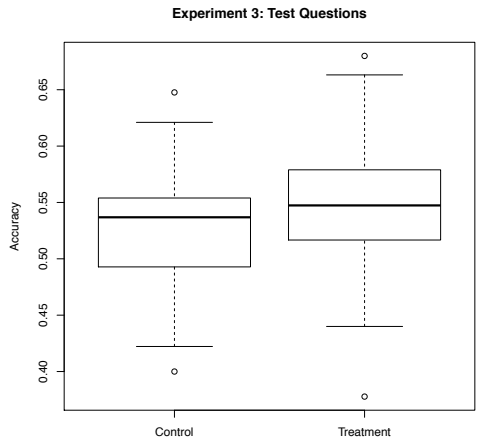

Figure 7: A boxplot of Accuracy across the different groups in Experiment 3 where test questions were used to guarantee a certain level of relevance judging accuracy.

both the control and treatment groups is much higher than it was for the previous experiments, as was to be expected, given that the poor performing crowd-workers (as judged by their accuracy on the test questions) were prevented from further relevance judging and their previous judgements removed from the analysis (if they had managed to judge less than 50 documents prior to being disqualified).

The spread of worker performance shown in Figure 7 is similar for the control and treatment groups, with the treatment group showing higher median (and mean) performance across crowd-workers, which is consistent with the hypothesis that providing a leaderboard motivates improved judging performance. However, the difference between the means was not found to be statistically significant at the 0.05 level with P-value of 0.071, indicating that a larger experiment was required. We therefore repeated the experiment three weeks later[13], doubling the sample size of crowd-workers involved. For the repeated experiment, the average accuracy was 52.9% for the 105 crowd-workers in the treatment group versus 51.2% for the 92 workers from the control group[14], and a significant difference in crowd-worker accuracy across the groups was observed with P-value of 0.039.

[13]We waited to repeat the experiment in order to reduce the chance that any crowd-workers from the previous experiment would participate in the repeated experiment.

[14]The fact that the macro-accuracy values (for both control and treatment groups) were lower for the repeated experiment than in the original experiment may be due to the fact that the same number (10) of $1 bonuses were offered twice the number crowd-workers.

4 Conclusions

In this paper we investigated the efficacy of providing real-time performance feedback to crowd-workers who are annotating document-query pairs with relevance judgements. The main findings of the investigation were:

1. Solely providing basic feedback to crowd-workers in the form of real-time performance information or a leaderboard has little effect on their relevance judging accuracy.

2. Providing a monetary incentive appears necessary in order to motivate crowd-workers to compete to achieve better annotation performance. Moreover, the provision of a leaderboard appears necessary in order to get the maximum effect of the payment of bonuses to the best performing relevance judges.[15]

3. Including test questions improves overall rating accuracy, and does not adversely affect the usefulness of a leaderboard.

There are a number of interesting directions for future work:

- We intend to repeat experiments for the real-life use-case where the ground truth relevance judgements are not known, and the accuracy of crowd-workers must be estimated based on the relevance judgements collected across workers up to that point. In this case, EM-based algorithms such as Dawid-Skene [DS79] can be used to estimate quality of each crowd-worker (as well as the posterior over relevance for each document). The estimated accuracy values will likely exhibit greater variability and it will be interesting to see whether that increased variability has an effect on crowd-workers actual accuracy.
- We intend to investigate other performance measures such as the amount of time taken to provide relevance judgements and the number of relevance judgements per worker to see whether real-time feedback and competition causes crowd-workers to annotate more items. We would also like to deepen the analysis of the collected data to investigate whether the accuracy of workers improves over time and what effect competition has on workers' performance over time.

[15]We note that the second finding was not significant at the 0.05 level, but the controlled manner in which we performed the tests, and the consistency of the improvement across experiments is indicative of the fact that larger repeat experiments would discover significant results.

- Finally, there is an enticing opportunity to provide real-time rewards to crowd-workers based on their performance and possibly even link the amount of bonus paid to the estimated accuracy of the crowd-worker, thereby leading to in some sense "economically optimal" (from a decision theoretic point of view) crowd-sourcing approaches.

Acknowledgments

The authors thank Wray Buntine for useful discussions regarding this work and Chunlei Chang for programming assistance. Mark Carman acknowledges research funding through a Collaborative Research Project award from National ICT Australia.

References

[ABY11] Omar Alonso and Ricardo Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *Advances in information retrieval*, pages 153–164. Springer, 2011.

[AM09] Omar Alonso and Stefano Mizzaro. Can we get rid of trec assessors? using mechanical turk for relevance assessment. *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 15–16, 2009.

[DS79] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):pp. 20–28, 1979.

[EHdVS12] Carsten Eickhoff, Christopher G Harris, Arjen P de Vries, and Padmini Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 871–880. ACM, 2012.

[GL10] Catherine Grady and Matthew 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, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Har14] Christopher G Harris. The beauty contest revisited: Measuring consensus rankings of relevance using a game. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, pages 17–21. ACM, 2014.

[KLSH09] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181, 2009.

[MN15] Cristina Ioana Muntean and Franco Maria Nardini. Gamification in information retrieval: State of the art, challenges and opportunities. In *Proceedings of the 6th Italian Information Retrieval Workshop, IIR'2015*, 2015.

[MPA15] Pavel Metrikov, Virgil Pavlu, and Javed A. Aslam. Aggregation of crowdsourced ordinal assessments and integration with learning to rank: A latent trait model. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1391–1400, New York, NY, USA, 2015. ACM.

[MWA+13] Pavel Metrikov, Jie Wu, Jesse Anderton, Virgil Pavlu, and Javed A. Aslam. A modification of lambdamart to handle noisy crowdsourced assessments. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, ICTIR '13, pages 31:133–31:134, New York, NY, USA, 2013. ACM.

[SKL12] Mark D Smucker, Gabriella Kazai, and Matthew Lease. Overview of the trec 2012 crowdsourcing track. Technical report, DTIC Document, 2012.