GeAnn at the TREC 2011 Crowdsourcing Track

Carsten Eickhoff
TU Delft
Netherlands
c.eickhoff@tudelft.nl

Christopher G. Harris
The University of Iowa
USA
christopher-harris@uiowa.edu

Padmini Srinivasan
The University of Iowa
USA
padmini-srinivasan@uiowa.edu

Arjen P. de Vries
CWI
Netherlands
arjen@acm.org

ABSTRACT

Relevance assessments of information retrieval results are often created by domain experts. This expertise is typically expensive in terms of money or personal effort. The TREC 2011 crowdsourcing track aims to evaluate different strategies of crowdsourcing relevance judgements. This work describes the joint participation of Delft University of Technology and The University of Iowa, using GeAnn, a term association game, we generate relevance judgements in an engaging way that encourages quality submissions, which otherwise would have to be motivated through rigid quality control mechanisms and additional incentives such as higher monetary rewards.

1. INTRODUCTION

Ground truth relevance assessments for information retrieval benchmarking initiatives such as TREC have traditionally been created by professionals with substantial expertise in search and information science. Typically, the assessment is done either in a collective effort of the research community by pooling and redistributing the submitted runs to participants, or, through external experts, such as the assessors at NIST [5]. Recent work has shown the applicability of crowdsourcing for this use case [1]. Under the right conditions, a group of inexpensive workers could match the performance of professional NIST annotators for this task at significantly lower cost. However, the novel setting introduces a number of new challenges, previously unknown in the traditional controlled relevance assessment task. Concretely, there are frequent mentions of crowdsourcing workers cheating or delivering results of inferior quality [3, 6]. The TREC 2011 crowdsourcing track was set up to devise and compare different strategies of how to phrase relevance assessment tasks as crowdsourcing HITs.

We suspect that there are fundamentally different motivations for offering workforce on a crowdsourcing platform [4]. *Money-driven* workers are mainly motivated by the financial reward that is being paid upon completion of the HIT. *Entertainment-driven* workers, on the other hand, primarily seek an appealing pastime while seeing the payment as a positive side effect rather than a central motivation. Due to these different underlying motivations, we expect to observe a different working behaviour. Financially driven workers may display a greater likelihood to cheat or take shortcuts that result in lower result quality. Our TREC participation aims at providing a more appealing and engaging relevance assessment environment by means of a term association game.

The remainder of this working note is structured as follows: Section 2 describes our game-based approach to the assessment task and gives a detailed inspection of the obtained results. Section 3 describes and evaluates the trust aggregation method used for the consensus task. Section 4 discusses general observations about the track and its setting. Finally, Section 5 closes by proposing future modifications of the GeAnn game and its use for relevance assessments.

2. ASSESSMENT TASK

The first task asked participants to collect binary relevance judgements for approximately 2100 topic/document pairs in typical TREC fashion. The effectiveness of the different strategies and HIT designs is evaluated in terms of the quality of the collected labels, the time taken to create those labels and the amount of money invested in the process.

2.1 Approach

A fundamental difference between typical relevance assessment tasks in the fashion of TREC and the game that we propose, resides in the fact that we do not judge the relevance of a document as a whole, but instead break down the global decision onto term level. In the game, the player is confronted with 4 buckets at the bottom of the screen, each of which represents a TREC query. From the top of the screen, a keyword (or image) slides down and the player is required to direct it into one of the query buckets that is most closely related to the term. One of the 4 buckets represents the original query from the q/d pair, 2 are randomly drawn TREC queries and a final bucket is labelled "None" to account for terms that are not related to any of the topics. The top left hand corner provides additional evidence by displaying the snippet of text in which the current term appeared on the original web page. Depending on the consensus with peers, the player is awarded points, that ultimately reward a position on the ladder board. As the game progresses, the speed increases, making decisions more difficult. The terms are aligned such on the screen, that without user input they will not fall into any of the buckets. As the player misses a bucket for the third time, the game ends. Figure 1 shows a screen view of the annotation game.

Consequently, we are faced with a number of preprocessing steps before being able to begin the annotation: (1) Break up the document into a set of sentences S. (2) Rank

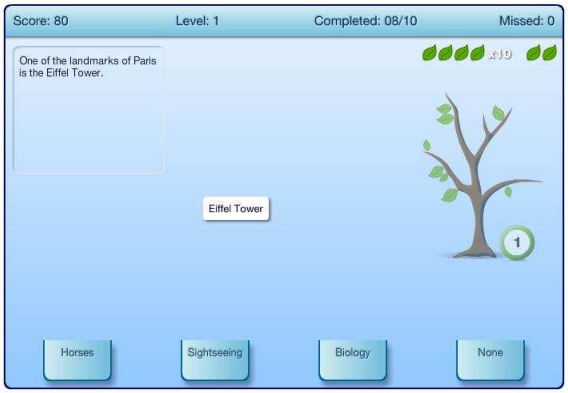

Figure 1: Screen view of the GeAnn game.

every sentence $s \in S$ according to an informativeness criterion $c(s)$. In the present case, we used the averaged idf score across all constituent terms $t \in s$. (3) Use the top n sentences from the ranked list for assessment by means of our game. (4) For each of the selected sentences, identify the single most informative term and use it as a sliding keyword, while the original sentence is shown as context snippet in the top left hand corner.

$$c(s) = \frac{1}{|s|} \sum_{t \in s} idf(t) \tag{1}$$

The main decision to influence the confidence in our page-wide relevance labels is the choice of n. High settings of this parameter result in a better coverage of the document's content in the assessment. To comply with the TREC deadline despite several delays in data distribution and game development, we were only able to judge a fixed number of $n = 6$ sentences per document. Optimally, concrete settings of n should depend on the document length, as well as the prior agreement rates on the document in question. In such a scheme, it would be possible to demand more assessments for ambiguous or long documents. Finally, in a post-processing step, we aggregate the sentence-level judgements that were made by individual players, and, subsequently, make a global decision l_{doc} across all sentences, games and players. The aggregation is based on a uniform majority voting scheme mv across a set of labels L, in which the most frequent label is propagated.

$$l_{(sent)} = mv(L_{game}) \tag{2}$$

$$l_{(doc)} = mv(L_{sent}) \tag{3}$$

The game was initialized in this fashion and the HITs were offered on Amazon Mechanical Turk via CrowdFlower. Workers were asked to play at least one round (10 terms from 2 document sets as provided by the TREC organizers) of the term association game and were offered a payment of 1 US cent. This was regarded to be more of an initial incentive for giving the game attention rather than an actual payment for the assessments. Moreover, to attract additional players, we advertised the game through various social networking sites.

Table 1: Game-based assessment behaviour

	All	Turk	Web
Games with 2+ rounds	52%	45%	58%
Rounds per game	6.7	5.3	7.9
Games per player	1.6	1.5	1.7
Returners	24%	20%	28%
Time to return	3.5h	3.4h	3.7h

2.2 Evaluation

In this section, we will inspect the performance of our game-based relevance assessment approach along the three previously mentioned dimensions: (1) Result quality, (2) Time taken to acquire results, as well as (3) the financial effort put into the assessment.

As mentioned before, we recruited players through word of mouth as well as an advertisement HIT with a very small payment. Throughout the evaluation section, we will pay careful attention to investigating whether there are significant differences in the observed assessment behaviour of paid vs. unpaid players. Overall, 47% of our 188 players that contributed to the TREC submission were recruited through the crowdsourcing HIT. The remaining players accessed the game directly from the Web. Table 1 presents an overview of a number of key statistics for the overall player population, as well as per subgroup.

Although no further payment was offered after the first round of 10 term associations in a game, we can see that a substantial number of players continued the game even without the prospect of an additional financial reward. The share of players recruited through MTurk to which this applies is slightly lower than for external players. The average game lasted between 5 and 8 rounds, with slightly shorter games being played by Turkers. On average, each player played 1.6 games, with shares of 20-28% of unique players returning for additional games after the first. For returning players, the average time between games was found to be 3.5 hours without major differences by player origin. The global distributions of rounds per game and games per player are shown in Figures 2 and 3, respectively. These initial figures already hint towards a tendency that is not typically found for crowdsourcing settings: Workers performing more work than they were required and paid for. A possible reason could be the engaging nature of our relevance assessment game. We will revisit this observation when analysing the cost efficiency of our method.

Label quality

One of the key performance criteria for relevance assessments is the accuracy of the collected labels. In our main TREC run, we evaluated the pure quality of the labels produced by our method. We omitted any form of aggregation or majority voting across players. The result can be considered a conservative lower bound performance that would be equivalent to asking relevance assessments to only a single worker without redundancy. The official TREC evaluation results with respect to the global consensus across teams as well as prior gold standard judgements from NIST are shown in Table 2. Even without any form of consensus, the use of which is typically considered mandatory for crowdsourcing, could we achieve substantial result quality. In relevance assessment scenarios, agreement rates of 60-70% are typically

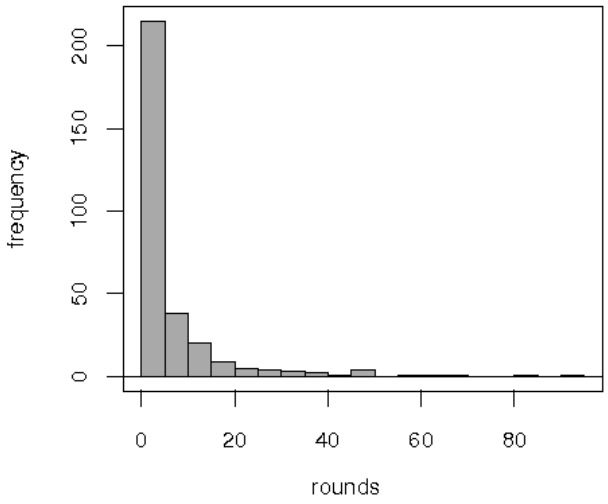

Figure 2: Distribution of rounds played per game.

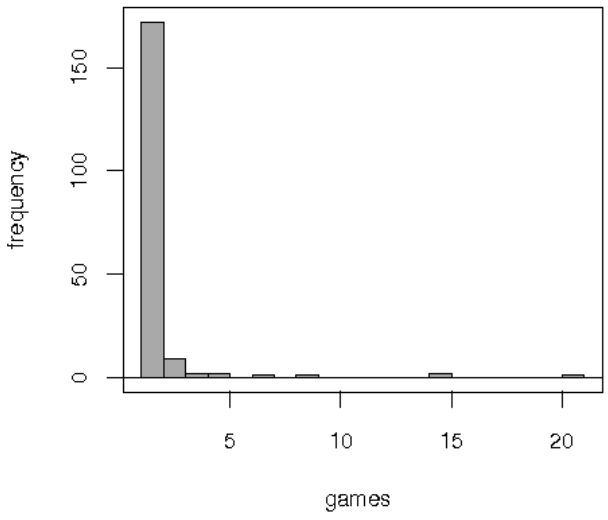

Figure 3: Distribution of games played per player.

Table 3: Aggregated Task 1 results

Source	Accuracy	R	P
Consensus	69.4%	79.7%	69.2%
Gold	63.1%	75.9%	73.7%

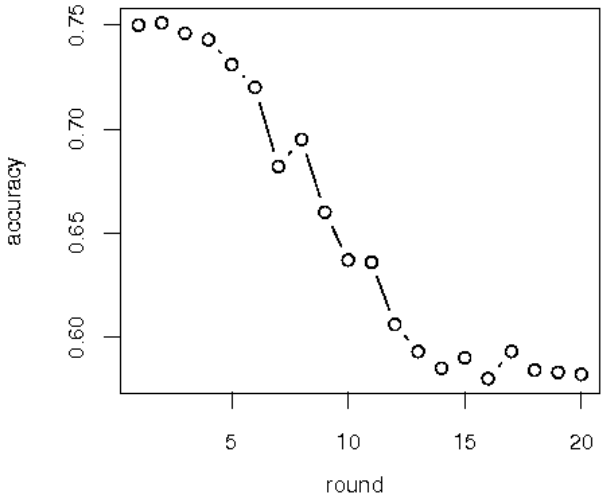

Figure 4: Label accuracy by game round.

to be expected from single judges in controlled lab environments. Being able to reach this quality level with single uncontrolled annotators was not to be expected.

To get more realistic insights into the potential of our method, we additionally aggregate majority vote labels across players. Table 3 reports the updated figures for this setting. We can note a consistent upwards tendency for all compared measures.

In order to create a competitive and challenging atmosphere that would motivate players to return to the game, we increase the game speed with each new round into which the player advances. This time pressure could have an influence on label quality as players have less time to make decisions in later rounds. Figure 4 shows the accuracy (agreement with global majority label) of judgements as a function of the game round in which they were issued.

Given this general downwards tendency in result quality, we reconsider and expand our majority voting scheme by a round-based confidence parameter λ. As the game speeds up, we expect players to err more frequently and therefore put less trust in judgements from higher rounds. Starting at $\lambda = 1.0$, for each round after the first, we reduce it by 0.05 to a minimum of $\lambda = 0.5$. Table 4 shows the resulting performance gains of this scheme.

Assessment speed

Another key criterion in the evaluation of crowdsourcing methods is the time required to produce a number of judgements. This global time between initially publishing the

Table 2: Official Task 1 results

Source	Accuracy	R	P	Specificity	Log Loss	KL Divergence	RMSE
Consensus	65.0%	76.9%	66.8%	45.4%	376.3	358.8	51.4%
Gold	62.3%	74.8%	72.4%	26.5%	94.6	94.6	52.8%

Table 4: Discounted Task 1 results

Source	Accuracy	R	P
Consensus	70.5%	80.1%	71.3%
Gold	64.3%	76.3%	74.8%

HIT and collecting the results is controlled by two central elements: (1) The uptake time t_{uptake} expresses the mean interval between workers starting new games. Depending on how appealing the HIT looks and how competitive the offered pay level is, the uptake time can vary greatly. Previous work has shown this factor to be subject to external influences such as the size of the HIT batch (in our case 50 HITS per batch) or its position on the overview page from which workers select their tasks. (2) The task time t_{task} represents the actual time a worker spends per task. A naive estimate of an upper bound on the expected runtime per batch is therefore:

$$t_{batch} = n(t_{uptake} + t_{tasl})$$

In reality, HITs are accessed by multiple workers in parallel, thus reducing the time per batch. Especially for short HITs with low values of t_{task}, the batch run time is dominated by the uptake rate.

In the concrete case of our TREC participation, we collected 10,535 document-level judgements within 8 days. During this period, we observed a t_{uptake} between games of 33 minutes. The overall distribution of games played is shown in Figure 5. The HIT uptake in the last 3 days is substantially lower as we were only issuing small batches by that time in order to fill in labels that had previously been begun in cancelled sessions. Due to the track's rules, in order to include a worker's judgement on a q/d pair, he has to judge all pair in a set. Therefore, we had to resubmit incomplete sessions.

The time per HIT in our case could not be evaluated statically as a game could run for a variable number of rounds. Instead, we measure the time taken per round of 10 judgements. The speed of the game imposes an upper bound on the available time per judgement, and, subsequently, per round. Concretely, we observed an average time between judgements of a round of 5.6 seconds, resulting in an average round duration of approximately one minute. As a conclusion to our temporal result analysis, we find our game environment to facilitate judgements in a fast, yet quality-preserving manner.

Assessment cost

The final part of our analysis is concerned with the necessary cost involved in the collection of crowdsourced relevance labels. In the previous sections, we found game-based relevance assessments to be of good quality and collection speed. The real strength of our method, however, lies in giving workers an alternative motivation from the pure financial reward. The total overall cost involved in the collection of our 10,350 query/document labels, including the AMT

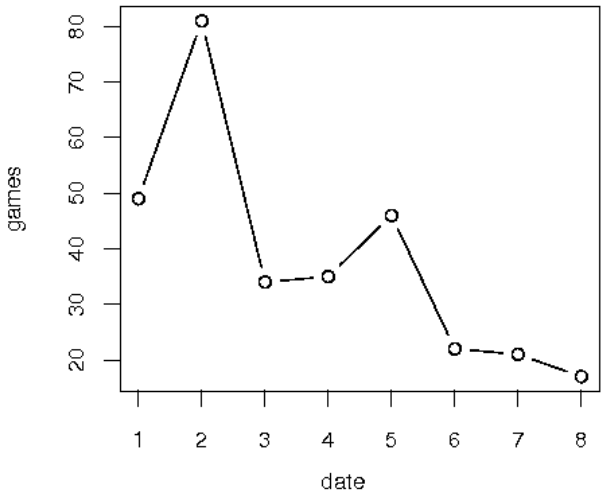

Figure 5: Distribution of games played over time.

service overhead, was $ 3.74. Even with respect to the generally low pay rates on crowdsourcing platforms, this result can be considered remarkable. It nicely shows how workers become players with a primary interest in the game experience rather than the hourly rate of only $ 0.23 (a total of $ 3.74 paid for 10350 labels, with the average assessment taking 5.6 seconds).

3. CONSENSUS TASK

The previous task was concerned with the collection of labels using crowdsourcing. Task 2 assumes that this step has already been taken. Given a number of crowdsourced relevance labels for query document pairs, determine consensus between multiple labels on the same pair. The collection contains 19,033 unique query/document pairs for which 89,624 binary relevance had been collected. 3,275 pairs also contain gold standard NIST labels. For the final evaluation, 1000 additional gold labels were withheld. The document IDs were anonymized so that no further evidence beyond the set of labels could be collected.

3.1 Approach

Our approach towards Task 2 is based on iterative computation of worker reliability in order to make non-uniform majority votes. A central component of our method is the reliability function r_t that assigns each worker w a score between 0 and 1 at time t. Higher scores express higher prior reliability. At the beginning of our iterative scheme, $r_{1.05}(w)$

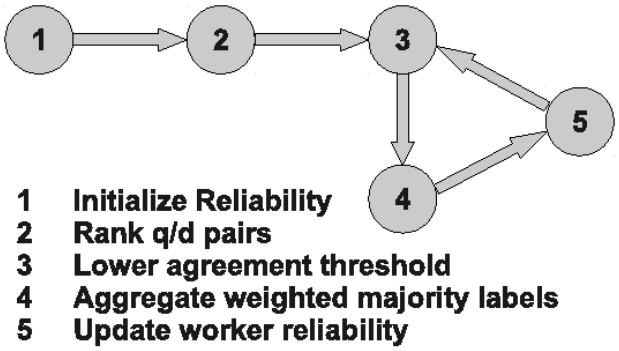

Figure 6: Reliability-based voting.

is initialized as the worker's accuracy $acc_g(w)$ on the set of gold labels G_w, that he encountered. If no gold judgements are available for that worker, we assume the maximum reliability score of 1.

$$r_{1.05}(w) = \begin{cases} acc_g(w) & if\ |G_w| > 0 \\ 1 & else \end{cases}$$

Now, we rank all query/document pairs in decreasing order of the ratio of agreement a in their labels. E.g., a pair for which all 5 workers assigned the same label ($a = 1$) would be ranked higher than one for which 2 relevant and 3 irrelevant judgements were issued ($a = \frac{2}{3}$). At this point, the preparation is complete. Now, we start the iteration process by computing majority voting labels $l_t(p)$ for all pairs p with an agreement $a >= t$, where t ranges from 1 to 0.5 in steps of 0.05 per iteration. The majority label, finally, is computed as the weighted average label, based on the individual worker labels $l(w, p)$ and the previous worker reliability.

$$l_t(p) = \frac{\sum_{w \in W} r_{t+0.05}(w) * l(w,p)}{\sum_{w \in W} r_{t+0.05}(w)}$$

The last step in each iteration is to update each worker's reliability scores by the accuracy of all previous votes (against both gold and consensus). Once this is achieved, we lower the threshold agreement and start the following iteration. Figure 6 illustrates the work flow of our method graphically.

$$r_t(w) = acc_t(w)$$

3.2 Evaluation

The official evaluation of Task 2 was conducted based on the overall consensus across groups as well as the 1000 held-out gold labels from NIST. Table 5 gives an overview of the achieved performance. Inspecting these numbers, we find, that while being able to aggregate worker performance, our method was not among the most competitive ones. We can additionally note a significant disparity between performance as evaluated against consensus and gold labels. This tendency is repeated for all participating groups. In a number of cases this drastically changes the ranking of teams between the two evaluation methods. We will discuss this property of employing consensus labels for evaluation purposes in greater detail in the following section.

4. DISCUSSION

Following the performance analysis of our proposed method, we will now proceed to discussing a number of central observations made during the label collection process.

Traditionally, especially in Web retrieval settings, relevance is assumed to be distributed sparsely in the document collection. For any reasonable query, we can expect the vast majority of documents to be irrelevant. Typically initiatives such as TREC resort to capturing a biased sample of retrieval results that gives a more even split between relevant and irrelevant pages than a random sample of the Web would. Especially for crowdsourcing purposes such an approach appears sensible in order to ward off workers who try to learn the underlying label distribution in order to cheat on subsequent tasks. The data set constructed for the crowdsourcing, however, makes an exception in the opposite direction, here, 68% of the provided NIST labels belong to the relevant classes. Also, the resulting crowdsourced consensus labels show a collection-wide average relevance of 0.55, that significantly surpasses an even split between classes. A global share of 57% of all labels belong to the relevant class. Figure 7 shows the distribution of relevance in the consensus labels across all teams. As we can see, the bias resides on the highly relevant pages which seem to be over-represented in the collection. Such an imbalanced setting bears significant dangers of over training the worker population towards giving relevance-biased answers.

This problem gains additional impact in consensus-based settings. As the majority across a collection of labels is used to evaluate subsets of the collection, one has to be very careful to avoid any form of crowd training where possible. As soon as large parts of the crowd suspect a biased underlying label distribution, global consensus may not represent a valid means of evaluation any more. This is additionally aggravated by the fact that majority voting across teams with homogeneous numbers of submissions is not necessarily an objective measure of result quality. Teams with a high number of submitted labels that were previously curated to follow an internal majority can greatly bias the global decision. As a consequence, the consensus label can be gamed by submitting more than other teams. To further understand whether this happened in the present evaluation, it would be good to investigate the correlation between the number of submitted labels and the various resulting team performance scores. Since these figures are not openly available, we could not provide the sketched analysis.

A final aspect to be covered by this discussion is crowd diversity. The submitted runs show a substantial variance in the number of unique workers who contributed to the submitted labels. The different teams employed worker pools ranging from 1 to 503 individuals with an average of 128 workers per team. Using a Spearman rank correlation test, we analysed team performance in dependency with the size of their worker pool. The result can be found in Table 6. We can notice a strong inverse correlation between the size of worker pools and the achieved performance in terms of accuracy and precision. While this observation does not necessarily imply a causality between small worker groups and superior performance, it certainly raises the question how comparable the employed settings were across teams and how well methods based on 30 workers scale to larger problems.

Table 5: Official Task 2 results

Source	Accuracy	R	P	Specificity	Log Loss	KL Divergence	RMSE
Consensus	73.6%	81.5%	78.0%	60.1%	5992.5	12911.1	15.2%
Gold	57.7%	73.5%	55.8%	41.8%	1150.4	1150.5	51.3%

Table 6: Team performance correlated to the size of employed worker pools

	Acc	P	R	# Workers
Acc	1.00	0.95	0.19	**-0.71**
P		1.00	-0.05	**-0.76**
R			1.00	-0.14
# Workers				1.00

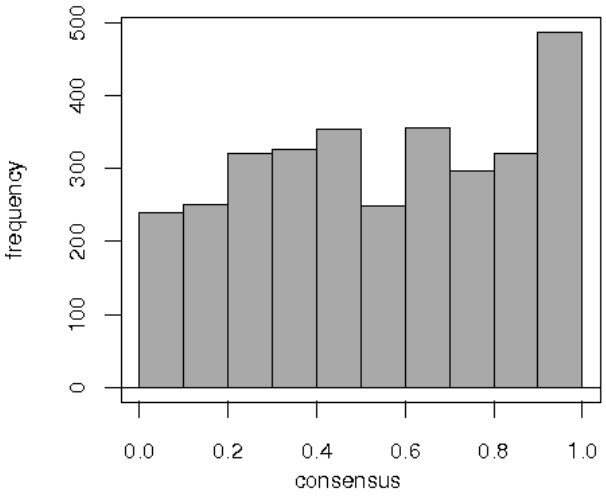

Figure 7: Distribution of relevance in consensus.

5. CONCLUSION & FUTURE DIRECTIONS

In this work, we described GeAnn, a game-based approach towards phrasing relevance judgement tasks in an engaging and entertaining way. By breaking up web documents into phrases and asking players to associate those with a number of topics we create relevance judgements in an efficient manner. The evaluation of the TREC 2011 crowdsourcing track has shown that our method delivers good quality at extremely low cost. Making annotations more entertaining served for an alternative motivation besides the financial reward. As a side-effect of our judgement scheme, we produce passage-level relevance judgements from which we derive a holistic decision per document. The main challenge resides in the fact that the task being carried out in our game (associating terms and topics) is not identical to the one being ultimately evaluated (page-wide relevance assessments between queries and documents). This disparity may introduce additional noise in the resulting judgements.

There are a number of aspects to be addressed and improved upon in the future in order to further improve the performance of game-based annotations. (1) Currently, there was no limit to the number of rounds for which a game could last. This confused players and shifted the aim of the game to "surviving" through as many rounds as possible instead of producing as high-quality labels as possible within a fixed number of rounds. An updated version of the game now has a fixed number of ten rounds after which the game ends. (2) currently, we extracted a fixed number of phrases per document. By doing so, we may over-represent short documents while having insufficient coverage of large documents. In the future, we will take a different approach that takes document lengths into consideration when extracting phrases. With respect to this, the ideal degree of document coverage has to be determined. (3) Currently, the only element of competition lies in ranking players on a ladder board. For subsequent versions of the game we would like to emphasize this point to further increase player engagement. This has been previously shown to be beneficial for result quality [2]. Concretely, we aim to introduce a multi player setting, in which the direct competition between peers will be enabled. (4) In this work, we exclusively focused on textual resources. However, images often convey a significant amount of meaning, as well. In the future we aim to also use images to replace some of the game elements (e.g., bucket labels, sliding terms or text blocks). This may introduce more variation in the game, thus additionally motivating players. We also suspect moving image content to be easier to discern than text (a few players commented on the sliding terms being hard to read).

We can conclude that there are many potential alleyways towards making game-based relevance assessments a superior alternative to both, standard expert assessments as well as crowdsourced tasks.

6. REFERENCES

[1] 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. Citeseer, 2009.

[2] M. Csikszentmihalyi. *Finding flow: The psychology of engagement with everyday life*. Basic Books, 1997.

[3] C. Eickhoff and A. de Vries. How crowdsourcable is your task? In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 11–14, 2011.

[4] C. Eickhoff, C. G. Harris, Srinivasan P., and de Vries A. P. Geann - games for engaging annotations. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, 2011.

[5] E. Voorhees, D.K. Harman, National Institute of Standards, and Technology (US). *TREC: Experiment and evaluation in information retrieval*. MIT press USA, 2005.

[6] J. Vuurens, A.P. de Vries, and C. Eickhoff. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, pages 21–26, 2011.