Quality through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments

Carsten Eickhoff
Delft University of Technology
Netherlands
c.eickhoff@tudelft.nl

Christopher G. Harris
Padmini Srinivasan
The University of Iowa
USA
{christopher-harris,padmini-srinivasan}@uiowa.edu

Arjen P. de Vries
Centrum Wiskunde & Informatica
Netherlands
arjen@acm.org

ABSTRACT

Crowdsourcing is a market of steadily-growing importance upon which both academia and industry increasingly rely. However, this market appears to be inherently infested with a significant share of malicious workers who try to maximise their profits through cheating or sloppiness. This serves to undermine the very merits crowdsourcing has come to represent. Based on previous experience as well as psychological insights, we propose the use of a game in order to attract and retain a larger share of reliable workers to frequently-requested crowdsourcing tasks such as relevance assessments and clustering.

In a large-scale comparative study conducted using recent TREC data, we investigate the performance of traditional HIT designs and a game-based alternative that is able to achieve high quality at significantly lower pay rates, facing fewer malicious submissions.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User / Machine Systems —*Human Factors*; H.1.2 [**Models and Principles**]: User / Machine Systems —*Human information processing*; H.5.2 [**Information Interfaces & Presentation**]: User Interfaces

Keywords

Crowdsourcing, Gamification, Serious Games, Relevance Assessments, Clustering

1. INTRODUCTION

In the course of the past 5 years, crowdsourcing has advanced from a niche phenomenon to becoming an accepted solution to a wide range of data acquisition challenges [10]. It has been used in the training and test phases of a great number of scientific projects and is firmly integrated into numerous evaluation and data acquisition schemes in academia and industry. One problem, however, seems to be inherent to the field; a significant share of the annotations created on crowdsourcing platforms are fraudsters' attempts to cheat the HIT (Human Intelligence Task) provider into paying them without having properly worked on the HIT. As a consequence, almost every scientific publication that employs crowdsourcing for data acquisition details the authors' tailor-made defense scheme against cheaters and a rising number of publications is exclusively dedicated to the task of detecting these individuals. The range of commonly-observed measures taken includes asking redundant questions to rely on an aggregate of several workers rather than the decisions of just one individual [33], using held-out data to compare with, asking plausibility questions and many more sophisticated methods [16, 13].

In many time-insensitive applications, HIT providers restrict the crowd of workers to certain nationalities (a typical example would be US workers only) who they trust will provide higher quality results. Although the approach is widely accepted and has been shown to significantly reduce the number of spam submissions, we believe that this uptake may be treating symptoms rather than the actual underlying cause. Rather than attributing the confirmed performance differences between the inhabitants of different countries to their nationality, we hypothesize that there are 2 major types of workers with fundamentally different motivations for offering their workforce on a crowdsourcing platform: (1) *Money-driven* workers are motivated by the financial reward that the HIT promises. (2) *Entertainment-driven* workers primarily seek diversion but readily accept the financial incentives as an additional stimulus. We are convinced that the affiliation (or proximity) to one of those fundamental worker types can have a significant impact on the amount of attention paid to the task at hand, and, subsequently, on the resulting annotation quality. We realize, that money-driven workers are by no means bound to deliver bad quality; however, they appear to be frequently tempted into sloppiness by the prospect of a higher time efficiency and therefore stronger satisfaction of their main motivation. Entertainment-driven workers, on the other hand, appear to work a HIT more faithfully and thoroughly and regard the financial reward as a welcome bonus. They typically do not indulge in simple, repetitive or boring tasks. We propose to more strongly focus on entertainment-driven work-

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.

ers by phrasing crowdsourcing problems in an entertaining and engaging way: As games. Csikszentmihalyi's theory of *flow* [12], a state of maximal immersion and concentration at which optimal intrinsic motivation, enjoyment and high task performance are achieved, further encouraged our design.

We intend to increase the degree of satisfaction entertainment-driven workers experience. This can lead to (a) higher result quality, (b) quicker batch processing rates, (c) lower overall cheater rates, (d) better cost efficiency. An additional incentive for delivering high-quality results in a game scenario would be the element of competition and social standing among players. Taking into account recent behavioural analyses of online communities and games [24], entertainment seekers can be expected to put considerable dedication into producing high-quality results to earn more points in a game to progress into higher difficulty levels or a rank on the high score leaderboard.

The novel contributions of this work are 5-fold: (1) We describe a game-based approach to collecting document relevance assessments in both theory and design. (2) Based on NIST-created TREC data, we conduct a large-scale comparative evaluation to determine the merit of the proposed method over state-of-the-art relevance assessment crowdsourcing paradigms. (3) Venturing beyond "hard" quality indicators such as precision, cost-efficiency or annotation speed, we discuss a wide range of socio-economical factors such as demographics and alternative incentives to enhance a fundamental understanding of worker motivation. (4) In a separate study, we demonstrate the generalizability of the proposed game to other tasks on the example of noisy image classification. (5) We create a corpus of relevance assessments of considerable size and quality that we disclose to the research community.

The remainder of this work is structured as follows: Section 2 describes the state of the art of crowdsourcing for document relevance assessments as well as in the domain of games with a purpose (GWAP). Section 3 introduces the theoretical considerations and design decisions that guide our annotation game. In Section 4, we describe the setup as well as the results of a large scale study conducted on several commercial crowdsourcing platforms in order to determine the usefulness of the proposed method. Section 5 demonstrates the generalizability of our method on the task of image classification. Section 6 is dedicated to a discussion of the key insights gained in the course of this work, before concluding in Section 7.

2. RELATED WORK

This section introduces related approaches from three different research areas at the intersection of which this work is situated, namely, relevance assessment, crowdsourcing and games with a purpose.

Document relevance assessments have been playing a central role in IR system design and evaluation since the early Cranfield experiments [42]. Explicit judgements of (the degree of) relevance between a document and a given topic are used as a proxy of user satisfaction. Based on such test collections, the results of retrieval systems can be compared without undergoing numerous iterations of user studies. As a leading actor in IR evaluation and benchmarking, NIST's Text Retrieval Conference (TREC) [43] looks back on two decades of assessing document relevance. In order to be suitable for the evaluation of state-of-the-art Web-scale systems, the requirements in terms of size, topical coverage, diversity and recency that the research community imposes on evaluation corpora have been steadily rising. As a consequence, the creation and curation of such resources becomes more expensive. To further ensure the scalability of test collection-based system evaluation, considerable effort has been invested into designing robust performance measures [6], selecting the right documents for evaluation [8] and inferring judgements from user interaction logs [19].

Crowdsourcing represents an alternative means of collecting and annotating large-scale data sets. By employing a large group of individuals, paid at transaction level, tasks of considerable size can be completed in a timely and affordable manner. Document relevance assessments have been shown to be a task that can reliably be fulfilled by crowd workers [4, 22, 15]. One of the fundamental challenges in crowdsourcing is overcoming malicious and sloppy submissions. Many effective schemes exist, ranging from aggregating the results of independent workers to the use of honey pot questions [17, 18, 14]. Marshall et al. discuss the importance of engaging HIT design on result quality [28]. Recently, several scientific workshops have been dedicated to pursuing how to use crowdsourcing effectively and efficiently [10, 26]. Most notably, TREC 2011 for the first time offered a dedicated crowdsourcing track [25] addressing the crowdsourced collection of document relevance assessments. In this work, we adopt the evaluation scheme and data set used there.

The majority of crowdsourced tasks are plain surveys, relevance assessments or data collection assignments that require human intelligence but very little creativity or skill. An advance into bringing together the communities of online games and crowdsourcing is being made by the platform Gambit [1], that lets players complete HITs in exchange for virtual currency in their online gaming world. This combination, however, does not change the nature of the actual HIT carried out, beyond the fact that the plain HIT form is embedded into a game environment. Instead, we propose using an actual game to leverage worker judgements.

A number of techniques have been designed to make participation in human computation efforts as engaging as possible. Perhaps the most effective technique among these is a genre of serious games called games with a purpose [38] which have been developed with the focus of efficient and entertaining transformation of research data collection into game mechanics. By equating player success in the game with providing quality inputs, the idea is to extract higher-quality data than is currently done with dull repetitive tasks such as surveys. More than half a billion people worldwide play online games for at least an hour a day – and 183 million in the US alone [29]. The average American, for example, has played 10,000 hours of video games by the age of 21 [31]. Channeling some of this human effort to gather data has shown considerable promise. People engage in these GWAPs for the enjoyment factor, not with the objective of performing work. Successful GWAPs include the ESP Game [37], which solicits meaningful, accurate image labels as the underlying objective; Peekaboom [41], which locates objects within images; Phetch [39], which annotates images with descriptive paragraphs; and Verbosity [40], which collects common-sense facts in order to train reasoning algorithms. The typical research objective of these GWAPs is to have two randomly-selected players individually assign mutually-

agreed document labels, with the game mechanics designed to reward uncommon labels. In contrast, the game mechanics of our proposed game is to encourage and reward consensus labeling. The Pagehunt game presented players with a web page and asked them to formulate a query that would retrieve the given page in the top ranks on a popular search engine to investigate the findability of web pages [27]. While task-specific games have been shown to be engaging means of harnessing the players' intelligence for a certain research goal, there has not been a formal investigation of the merits of game-based HITs over conventional ones. Additionally, current GWAPs are typically highly tailored towards a certain (often niche) problem at hand and do not lend themselves for application across domains. We will demonstrate the generalizability of our approach in Section 5.

3. METHODOLOGY

In this section, we will introduce the annotation game as well as the necessary pre- and post-processing steps in order to acquire standard topic/document relevance assessments from it. Careful attention will be paid to highlighting motivational aspects that aim to replace HIT payment by entertainment as a central incentive.

3.1 Game Design

The central concept of our proposed game is to require players to relate items to each other. In order to preserve its general applicability, we tried to make as few as possible assumptions about the nature of those items. The game shows $n = 4$ concept buckets $b_1 \dots b_n$ at the bottom of the screen. From the top, a single item i slides to the bottom, and has to be directed into one of the buckets by the player. Doing so expresses a relationship between i and b_j. Additional information about i can be found in an info box in the top left corner.

In the case of document relevance assessments, the concept buckets b_j display topic titles, item i is a keyword from a document and the info box displays the context in which the keyword appears in that document. Figure 3.1 shows a screenshot of our game. A live version can be found online[1]. For each assigned relation between i and b_j the player is awarded a number of points. The score in points is based on the degree of agreement with other players. In addition to this scheme, the game shows a tree that grows a leaf for every judgement that consents with the majority decision. A full tree awards bonus points. In this way, we reward continuous attention to the game and the task. As a final element, the game is divided into rounds of 10 judgements each. After each round, the speed with which item i moves is increased, making the task more challenging to create additional motivation for paying close attention.

After up to 5 rounds (the player can leave the game at any point in time before that) the game ends and the achieved points as well as the player's position in our highscore leaderboard are shown. Together with the log-in concept, this aims to encourage replaying as people want to advance into the higher ranks of the leaderboard.

3.2 Data Pre-processing

Previously, we described the mechanics and underlying assumptions of the game. Now, we will detail how to set

[1] http://www.geann.org

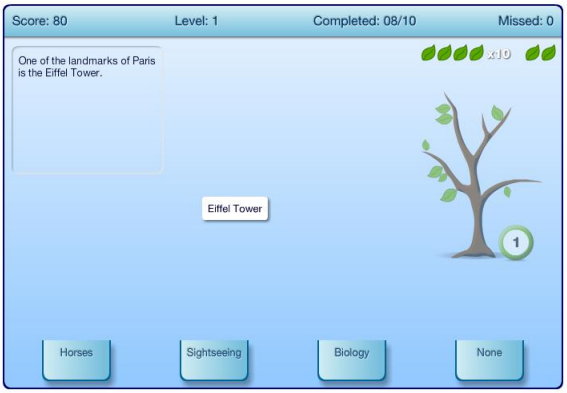

Figure 1: A screenshot of the annotation game in which the keyword "Eiffel Tower" has to be related to one of a number of concepts.

it up with TREC data for relevance assessments. All relevant game information is stored in a relational database from which items, concepts and additional information are drawn and into which user judgements are stored, subsequently. We assume a list of pairs p consisting of a query q and a document d for which we will collect user judgements. For every d, we extract the textual web page content and break it up into a set S of sentences $s_{d,1} \dots s_{d,|S|}$ using the LingPipe library [5]. In the following step, S is reordered by a ranking function $r(s)$, based on decreasing mean inverse document frequency (idf) of sentences s with a number of $|s|$ constituent terms t_n. $|C|$ denotes the number of documents in the collection and $df(t)$ is the number of documents containing term t. In this way, we promote the most salient and informative sentences in the document. The underlying $idf(t)$ statistics for this step are computed collection-wide.

$$idf(t) = \frac{|C|}{df(t) + 1}$$

$$r(s) = \frac{1}{|s|} \sum_{n=1}^{|s|} idf(t_n)$$

Finally, the k highest-ranking sentences (those with the highest scores of $r(s)$) are selected and used in the game. The concrete setting of k depends on the size of document d and was set to $0.1|S|$ in order to account for different document lengths. Higher settings of k result in a higher judgement density per document. For each of the selected sentences, we extract the single highest-idf term t_n as sliding keyword i, and the full sentence as context information to be shown in the top left corner. The concept buckets b include the original query q, two randomly selected topics from the database, as well as an "other" option to account for none of the offered concepts being related to i. The buckets are shown in random order to prevent selection biases.

3.3 Assessment Aggregation

As a final step, we have to transform the players' conceptual associations into document-wide relevance judgements. Each player annotation a can be understood as a quintuple $a = (p_a, i_a, s_a, c_a, r_a)$ in which player p associated item i occurring in context sentence s with concept c in round r of the game.

In a first step, we map all associations a to relevance votes. We interpret associations of any $s \in d$ to the concept of the original query q as a player's binary relevance vote $v_{p,s,q,r}$ between sentence s and query q as described in Equation 1.

$$v_{p,s,q,r} = \begin{cases} 1 & \text{if } c_a = q \\ 0 & \text{else} \end{cases} \quad (1)$$

In order to account for wrong associations and diversity in personal preference, we aggregate a global sentence-level vote $v_{s,q}$ across all players p. As the game speeds up in higher rounds, players have less time available for making the relevance decision. In a preliminary inspection of annotation results, we noticed significant drops of accuracy across subsequent rounds of the game. In order to account for this effect, we introduce a weighting parameter λ_r representing the confidence that we put into judgements originating from round r of the game being correct. For simplicity's sake, we reduce the confidence by 0.05 per round after the first one. Alternative strategies could for example include learning this parameter as a maximum likelihood estimate across previous observations. Equation 2 details the aggregation to a global sentence-level vote $v_{s,q}$ across the set of players $P_{s,q}$ that had encountered the combination of sentence s and query q.

$$v_{s,q} = \frac{1}{|P_{s,q}|} \sum_{p_i \in P_{s,q}} \lambda_r v_{p_i,s,q,r} \quad (2)$$

Finally, we aggregate across all sentence-level votes $v_{s,q}$ of a document d in order to get one global page-wide judgement that is comparable to well-known (e.g., NIST-created) annotations. Equation 3 outlines this process formally. It should be noted, that omission of this third step may, given the application at hand, be beneficial for the evaluation of tasks such as passage-level retrieval or automatic document summarization.

$$v_{d,q} = \frac{1}{|D|} \sum_{s_i \in D} v_{s_i,q} \quad (3)$$

4. EXPERIMENTATION

4.1 Research Directions

In this section, we describe the setup and results of a large-scale experiment conducted on several major commercial crowdsourcing platforms. Our performance comparison of traditional and game-based HITs will be guided by the following 7 fundamental directions:

Quality. How does the result quality of game-based crowdsourcing HITs relate to that of traditional ones given the same underlying crowd of workers and comparable financial means? We evaluate result quality in terms of agreement with gold standard NIST labels as well as with consensus labels across all participanting groups of the TREC 2011 Crowdsourcing Track [25].

Efficiency. Are game-based HITs more popular, resulting in quicker HIT uptake and completion than conventional ones? We investigate time to completion (for a given batch size) as well as the duration per document and per single vote.

Incentives. How much is fun worth? We investigate the influence of incentives such as fun & social prestige vs. monetary rewards on the uptake rate of HITs. Do users prefer entertaining HIT versions even though they pay less?

Consistency. Does our game encourage a stronger task focus, resulting in better within-annotator consistency? We investigate this dimension using a fixed set of workers (of varying reliability and quality levels) who are exposed to re-occurring HIT questions in a game-based or conventional setting to measure their self-agreement as an estimate of consistency and alertness.

Robustness. Does the share of (alleged) cheaters attracted to our game / attracted to conventional HITs differ? Independent of the overall result quality, the observed cheater rate is a surrogate of how reliable results are and how much sophistication a HIT should dedicate to fraud protection.

Population. Does the use of games lead to a different crowd composition? Offering game-based and conventional HITs, we collect surveys to investigate whether game-based HITs attract different kinds of workers.

Location. State-of-the-art crowdsourcing approaches frequently filter their crowd by nationality in order to improve result quality. We investigate whether there are indeed geographical preferences for game-based or conventional HITs and whether those can be related to the crowd composition in those areas.

4.2 Experimental Setup

In this comparative study, we replicate the setting that was proposed in the TREC 2011 Crowdsourcing Track assessment task [25]. A total of 3200 topic/document pairs (30 distinct topics, 3195 unique documents) were judged for relevance. The documents are part of the ClueWeb09 collection [7], and the topics originate from the TREC 2009 Million Query Track [9]. A comprehensive list of all topics and document identifiers are available from the TREC 2011 Crowdsourcing Track home page[2]. We contrasted the performance and characteristics of our proposed gamified HIT (Section 4.2.2) with those of a traditional one (Section 4.2.1). To attribute for assessment mistakes and personal preference, we collected judgements from at least 3 individual workers per topic/document pair in both settings. All HITs were run in temporal isolation (No more than 1 batch at any given time) to limit mutual effects between the tasks. In the following, we describe the respective task designs in detail.

4.2.1 Traditional HIT

As a performance baseline, we designed a state-of-the-art relevance assessment HIT. Its design follows accepted insights from previous work as detailed in the following. In order to limit the number of context changes, the document is shown in-line on the platform's HIT form as proposed by Kazai [20]. In this way, no distracting opening and closing of windows or browser tabs is required. To further enhance the task, we highlight every occurrence of query terms in

[2] https://sites.google.com/site/treccrowd2011/

the document. This technique was reported to be beneficial by several previous approaches, e.g., [36]. Finally, in order to deal with malicious submissions, we measure agreement with NIST gold standard pairs of known relevance. Workers who disagree on more than 50% of the gold labels are rejected from the judgement pool. In the HIT instructions, we briefly introduce the available relevance categories. The definition of relevance was introduced according to the TREC guidelines [25]. The HIT form contains 2 questions:

```
1. Please indicate the relevance of the shown
document towards the topic "<T>".

2. Do you have any remarks, ideas or general
feedback regarding this HIT that you would like to
share?
```

For each HIT, the place holder <T> is replaced by the current topic. Offering the possibility for worker feedback has been frequently reported to improve task quality and track down bugs or design flaws quickly [3]. The HIT was offered at a pay rate of 2 US cents per topic/document pair assessments; a reward level previously found adequate given the task [2].

4.2.2 Gamified HIT

The central piece of our proposed gamified version of the relevance assessment HIT is the annotation game that was described in Section 3.1. Instead of having the workers complete tasks locally on the crowdsourcing platform, the technical requirements of our game demanded running it off-site on a dedicated server. In order to verify task completion, workers are provided with a confirmation token after playing one round of the game (10 term associations). Back on the crowdsourcing platform, they enter this token in order to get paid. As a consequence, the actual HIT contained only a brief instruction to the off-site process and two input fields:

```
1. Please enter the confirmation token you
obtained after completing one round of the game.

2. Do you have any remarks, ideas or general
feedback regarding this HIT that you would like to
share?
```

Again, we solicit worker feedback. The HIT was offered at a pay rate of 2 US cents for one round (10 term associations) of the game.

4.3 Evaluation

All experiments described in this section were conducted between December 2011 and February 2012 on two crowdsourcing platforms: Amazon Mechanical Turk [35] as well as all available channels on CrowdFlower [11]. Initial evaluation did not show any significant differences in the work delivered by workers from different platforms. We will therefore not split the pool of submissions along this dimension. In total, 795 unique workers created 105,221 relevance judgements via our game. Additionally, 3000 traditional relevance judgements were collected for comparison. In total, we invested $90 to collect a volume of 108,221 annotations across the two compared experimental conditions. Together with the TREC 2011 Crowdsourcing Track annotations and the original NIST labels, this makes the T11Crowd subset of ClueWeb09 one of the most densely-annotated Web resources known to us. To enable reproducibility of our insights and to further general crowdsourcing research, the complete set of our judgements and the game itself are available to the research community[3].

Table 1: Annotation quality.

HIT type	Accuracy (NIST)	Accuracy (TREC-CS)
Conventional	0.73	0.74
TREC-CS	0.79	1.0
Game (plain)	0.65	0.75
Game (sent)	0.77	0.87
Game (doc)	**0.82**	**0.93**

Table 2: Annotation quality as a function of the game round in which judgements were issued.

Round	Accuracy (NIST)	Accuracy (TREC-CS)
1	0.72	0.81
2	0.67	0.77
3	0.62	0.73
4	0.60	0.69
5	0.54	0.65

4.3.1 Quality

As a starting point to our performance evaluation of game-based crowdsourcing of relevance assessments, we investigate the quality of the collected labels. Table 1 details the performance of our game in terms of overlap with gold standard NIST labels as well as the global consensus across all TREC 2011 Crowdsourcing Track participants (TREC-CS). We can note that already the conventional HIT delivers high result quality. Ratios between 65% and 75% are often considered good rules-of-thumb for the expected agreement of faithful human judges given a relevance assessment task [44]. TREC consensus labels show a high overlap with NIST annotator decisions. The third row in Table 1 shows the performance of direct unaggregated sentence-level votes from our game as described in Equation 1. While agreement with the TREC crowd is already substantial, the overlap with high-quality NIST labels lags behind. As we aggregate across multiple workers' annotations of the same sentence (Equation 2) and, finally, across all sentences extracted from the same document (Equation 3), the performance rises significantly, outperforming all compared methods. We used a Wilcoxon signed rank test at $\alpha < 0.05$-level to test significance of results.

In order to confirm the usefulness of our assumption from Section 3 concerning the decline of label quality as the game speeds up and the player has less time to make decisions, we evaluated annotation performance of raw labels according to the round in which they were issued. Table 2 shows a near-linear decline in agreement of plain game scores with TREC consensus as the game progresses. Agreement with NIST scores also consistently shrinks from round to round.

Finally, previous work on prediction in crowdsourcing systems demonstrates that reliability of the average predicted scores by the crowd improves as the size of the crowd in-

[3] http://sourceforge.net/projects/geann/

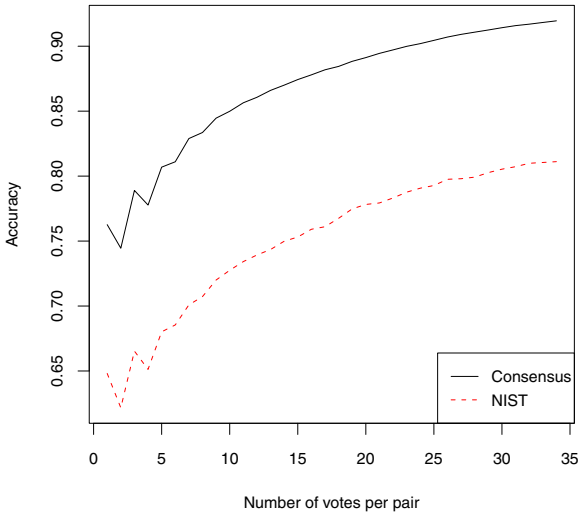

Figure 2: Quality as a function of votes per pair.

creases [34, 30]. The benefit of our game-based HIT is its popularity that allows us to collect more judgements per topic / document pair than traditional HITs. On average, each topic / document pair in the collection received 32 unique user judgements at the sentence level (some of which may originate from the same user as she or he rates different passages of the same document). Figure 2 shows how annotation quality develops as we add more judgements per pair. After initial fluctuation, as single votes have great influence on the overall decision, accuracy consistently improves as we add more votes per pair. The effect levels out as we approach the upper performance limit.

4.3.2 *Efficiency*

The second official performance indicator besides label quality in the TREC 2011 Crowdsourcing Track was the time necessary to collect the required judgements. For many use cases in human computation, low latencies are essential. The particular nature of the game imposed a time limit on players within which they had to make their decisions. As we detailed in the previous section, aggregation across users, weighting votes according to the difficulty level under which they were created, ensured competitive result quality. At the same time, however, a sequence of, individually quick, concept matchings enables workers to be more efficient and motivated than in conventional settings. Table 3 shows how conventional HITs take slightly longer to judge documents even when aggregating the duration of all passage-level votes in the game-based setting. Taking into account the significantly higher uptake rate (number of judgements issued per hour) of the game HITs, this serves for a considerably more efficient batch processing.

Especially in conjunction with the previous section's findings of high degrees of redundancy serving for better result quality, high uptake rates become crucial as they allow for timely, yet accurate decisions.

Table 3: Annotation efficiency.

	Conventional	Game-based
t per vote	40.1 sec	5.2 sec
t per doc	40.1 sec	27.8 sec
Uptake (votes per hour)	95.2	352.1

Table 4: Game-based assessment behaviour.

Criterion	Observation
Games with 2+ rounds	70.9%
Rounds per game	3.5
Players with 2+ games	79.5%
Games per player	4.36
Time between games	7.4 hrs

4.3.3 *Incentives*

The third and final evaluation criterion employed for TREC 2011 was the cost involved in the collection of relevance labels. With our game-based approach, we aim to, at least partially, replace the financial reward of the HIT with entertainment as an alternative motivation. In this section, we will investigate to which degree this change in incentives can be observed in worker behaviour.

In order to be paid via the crowdsourcing platform, workers had to complete at least one round (10 concept matchings) of our game. At that point the required confirmation token was displayed to them and they could return to the platform in order to claim their payment. However, the game offered an additional 4 levels to be played. From a purely monetary-driven perspective there would be no reason for continuing to play at that point. As we can see in Table 4, however, over 70% of games are played beyond the first round. This essentially results in crowdsourcing workers creating judgements free of charge because they enjoy the game experience. Additionally, we can observe players to return to the game after a number of hours to play again and improve their score and their resulting position on the leader board. Subsequent visits often happen directly to the game page, without being redirected from (and paid through) the crowdsourcing platform. Almost 80% of all players (633 out of 795) return after their first round played, with an average time gap of 7.4 hours between games. For regular HITs, we observed a return rate of only 23%.

When inspecting the concrete distribution of judgements across workers, as shown in Figure 3, we see this trend continued. Crowdsourcing tasks often tend to exhibit Power-law distributions of work over unique workers with some strong performers and a long tail of casual workers who only submit single HITs. Here, however, we notice a strong center group of medium-frequency players. We hypothesise that replacing the workers' extrinsic motivation ("*do the HIT to earn money*") by an intrinsic one ("*let's have some fun*"), causes these tendencies.

This has a number of noteworthy consequences: (1) We can attract workers to a HIT at a comparatively low pay rate. Even without playing beyond the first round, 2 US cents for 10 concept associations would roughly result in a prospective hourly pay of $1.20. (2) Furthermore, as most workers continue playing, additional annotations are created with no expectation of financial compensation. (3) Drawn by the competitive aspect of the game, workers re-

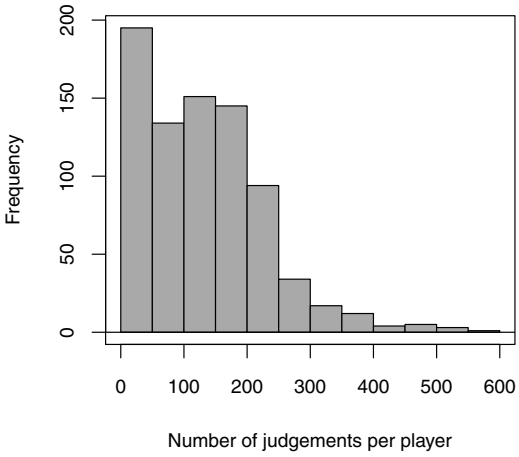

Figure 3: Distribution of judgements across users.

Table 5: Effective annotation cost.

	Conventional	Game-based
Cost per doc	$0.06	$0.0004
Whole corpus	$192	$1.28
Effective hourly rate	$1.80	$0.18

turn and create even more unpaid assessments. As a consequence, the overall amount of money invested into the game-based collection of more than 100,000 sentence-level relevance judgements was $27.74. This includes all administrative fees charged by the crowdsourcing platforms. In comparison, the participants to the TREC 2011 Crowdsourcing Track reported overall costs of $50 - $100 for the collection of significantly fewer labels.

Table 5 shows a final cost comparison of the conventional and game-based versions of the inspected relevance assessment HIT. While all indicators of cost efficiency from a worker's perspective clearly speak for choosing the conventional, better-paying HIT, the previously described figures of HIT uptake rates as well as the high number of alternative HITs available at all times on large-scale platforms such as AMT, indicate, that we reach workers who consider the entertainment potential of a HIT before choosing it. If we consider all judgements made in rounds after the first one and all judgements from revisits that were not paid for on the crowdsourcing platform as free-of-charge judgements, we arrive at a share of 83.7% of all labels having been created free of charge. Additionally, a number of players (39 out of 795) accessed the game without being prompted (and paid) by a crowdsourcing platform. These players were recruited from the authors' professional and private networks or word of mouth of other players. We could not find significant differences in the judgement quality or volume created by this group of players. The invested amount of money can be seen as advertisement costs rather than actual payments. In a traditional setting, collecting the same annotation density would have cost $2104.

4.3.4 Consistency

Following Csikszentmihalyi's theory of Flow [12], a state of deep immersion is a good foundation for high performance independent of the concrete task at hand. With our game-based HIT, we aimed to exploit this observation in order to create greater task focus than workers typically achieve on conventional HIT types. The previously shown result quality figures support this hypothesis. As an additional performance indicator, we will measure the workers judgement consistency. Faced with the same passage of text and choice of concepts multiple times, a situation-aware worker is expected to display a high degree of intra-annotator agreement. In the course of our judgement collection, we showed identical assignments to workers 837 times and observed an intra-annotator agreement of 69.8%. We set up a dedicated crowdsourcing experiment in which a portion of the offered topic / document pairs re-occurred. The HIT was set up following the scheme described in Section 4.2.1. Across 500 redundantly issued assignments, we observed an intra-annotator agreement of only 61.3%, a significantly lower ratio (determined using Wilcoxon signed rank test at $\alpha < 0.05$) than in the game-based setting. While the game setting resulted in higher consistency than usual crowdsourcing schemes, we could not match the consistency Scholer et al. [32] report for professional assessors as for example employed by NIST.

4.3.5 Robustness

Cheating, spamming and low-quality submissions are well-known and frequently-observed incidents on commercial crowdsourcing platforms. Previously, we demonstrated convincing result quality of gamified document relevance assessments when labels are aggregated across a sufficiently large number of workers. Since our approach appeals more to the entertainment-seeking rather than money-driven workers, we did not include a dedicated cheat detection scheme as would often be considered necessary in state-of-the-art HITs. However, we realise that the observed cheat rate in an assignment can serve as a surrogate for the confidence and reliability of the overall results. To this end, we measure the observed proportion of cheat submissions to our game as well as to the conventional HIT version. Eickhoff et al. [14] suggest categorizing workers who disagree with the majority decision in more than half of all cases as cheaters. In order to deliver a conservative estimate of the rate of cheat submissions, we tighten their definition and consider a worker as cheating if at least 67% of their submissions disagree with the majority vote. This scheme was applied to both, the conventional HIT as well as the gamified version. In the game-based case, we additionally flagged all submissions as cheat that tried using forged confirmation tokens. Overall, this resulted in a share of 13.5% of the conventional HIT's judgements being considered cheated. For the game-based version, the percentage was a significantly lower 2.3%. This finding conforms with [13], who observed innovative, creative tasks being less likely to be cheated on.

4.3.6 Population

In this work, we did not make use of any form of *a priori* filtering the pool of workers eligible to access our HITs. We hypothesise, however, that HIT type, financial reward and task phrasing influence the underlying crowd that decides to work on a given assignment. To better understand the com-

Table 6: Composition of the crowd

	Conventional	Game-based
Preference	39%	37%
Female	47%	35%
Age	34	27
Univ. degree	46%	62%
Income	$20k	$45k
English Native Speaker	24%	25%

position of the group of commercial crowdsourcing workers that are interested in games, we accompanied parts of our HITs by surveys in which we asked for high-level participant demographics and their preference for either the conventional or the game-based HIT. Table 6 shows an overview of several salient outcomes of the survey. The split in decisions was roughly equal, with 24% of workers not indicating clear preferences. The entertainment-seeking worker is on average several years younger, more likely to hold a university degree and will typically earn a higher salary. Finally, women were found to be significantly less interested in games than their male co-workers. This conforms with general observations about gender differences made for example by [23]. A worker's language background did not influence his or her likelihood to prefer games.

4.3.7 Location

Many commercial crowdsourcing schemes report performance gains when filtering the crowd by nationality. Due to different expected levels of education, language skills or cultural properties, such steps may influence result quality. As a final dimension of our investigation of games for use on commercial crowdsourcing platforms, we will inspect whether worker origin has an influence on result quality. From our survey, we found Indian workers, with a share of 60%, to be the dominant group in both settings. US workers were consistently the runners-up with a proportion of approximately 25%. There was no significant difference in the likelihood to prefer games over conventional HITs between countries.

Finally, when inspecting result quality from our game, again, no difference in performance or likelihood to cheat could be found. This suggests that filtering workers by nationality may not be ideal. In fact, the underlying worker motivation and HIT type preference may be assumed to have a far greater impact on observed uptake, performance and trustworthiness.

5. IMAGE CLASSIFICATION

In the previous sections, we described and evaluated the performance of the proposed crowdsourcing-powered annotation game for the task of TREC-style document relevance assessments. To demonstrate the generalization potential of the described concept-matching method, we applied the same game in an image classification pilot.

In the course of the Fish4Knowledge project (`http://www.fish4knowledge.eu/`), several underwater cameras have been placed in selected locations in south-east Asian coral reefs. The continuous recordings are supposed to further knowledge about behaviour, frequency and migration patterns of the resident tropical fish species. A key step to coping with the large amounts of image data produced by these cameras is a reliable automatic species classification. In order to train such systems, numerous training examples are required. While the project employs a team of marine biologists, their greater expertise is costly. Using our annotation game, we crowdsource the task of classifying the encountered species. Instead of relating keywords to TREC topics, the objective is now to match a shot of the underwater camera (often low quality) to high-quality examples of resident fish species. By initializing the underlying database with images rather than textual items, no changes to the actual game were necessary. Figure 4 shows a screenshot of this alternative game setting.

Figure 4: GeAnn applied for image grouping.

Our feasibility study encompassed 190 unique underwater camera shots for which known gold standard labels created by the marine biologists existed. Each biologist had classified all images, allowing us to contrast crowd agreement with expert agreement. The HIT was offered in January and February 2012 at the same pay rate (2 US cents per round of 10 associations) as the text-based version. Table 7 shows the results of the experiment in which the degree of agreement with the majority of experts as well as the crowd's inter-annotator agreement are detailed. We can see high agreement with expert labels as well as substantial agreement among workers. The popularity (qualitatively perceived through worker feedback) and uptake rate of this HIT even slightly exceeded those of the game-based one for document relevance assessments. Several workers had mentioned difficulties reading the moving text fragments in the short time available. With images, this does not appear to be an issue.

Methods like this could play an essential role either in the creation of training and evaluation data necessary for the assessment of automatic classification quality, or as part of a hybrid human-computer classification in which automatic methods narrow down the range of potential species before human annotators select the most likely species from the pre-selection. It should be noted, however, that the domain experts are by no means obsolete in this setting. While they annotated fish images simply based on their knowledge of the resident species, players of our game only had to select one out of a range of 4 species by similarity.

6. DISCUSSION

In the previous section, we found convincing results across all inspected performance dimensions, supporting the bene-

Table 7: Image classification performance

	Agreement (exp.)	Inter-annot. Agreement
Experts	0.82	-
Game	0.75	0.68

fit of offering alternative incentives besides the pure financial reward. In this section, we discuss a number of observations and insights that were not yet fully covered by the evaluation.

Firstly, considering the fact that the round concept of the game appears to invite workers to create assessments without payment (by playing on after having received the confirmation token), it is not obvious why we should limit the game to a fixed number of rounds. In the present setting, a game inevitably ends after the fifth round. One might argue that a higher number of rounds or even an open-ended concept would result in even greater cost efficiency. In fact, the opposite seems to be the case. In an initial version of the game, there was no upper limit to the number of rounds per game. As a consequence, some players were frustrated, as the only way to "finish" the game would be to either lose or give up. This resulted in fewer returning players. Additionally, the quality of annotations resulting from higher rounds was highly arguable as the objective of the game became mainly surviving through as many rounds of fast-dropping items as possible, rather than making sensible assessments. In the new, limited, setting, the clear objective is to do as well as possible in 5 rounds. Players who want to improve their score beyond this point have to return and start a new game.

A second key observation to be made is the fact that while we evaluate against the performance of NIST assessors and TREC participants, the tasks our workers face is a significantly different one. In the game, no worker gets to see full textual documents or is even told that the objective is to determine topical relevance of web resources towards given information needs. We deliberately aimed for such a loose coupling between game and task as we wanted to keep the game experience entertaining without the "aftertaste" of working. It is interesting that mere conceptual matching correlates well with actual relevance assessments. Also, in the data pre-processing phase, we do not extract sentences based on query terms but rather focus on pure *idf* figures as described in Section 3.2. In this way, we manage to capture the general gist of a document without artificially biasing it towards the topic.

Finally, the key insight gained from this work was the substantial benefit achieved by offering an alternative incentive to workers. Most of the interesting properties observed in the gamified system, such as workers producing free labels, would not have happened otherwise. This is, however, not necessarily limited to gamified tasks. In this paper we used games as one possible means of showing how a particular incentive (money) can be replaced with another one (entertainment). By doing so, we focus on a certain type of worker, entertainment-seekers, the existence of which we hypothesised based on previous experience with crowdsourcing. We are convinced that a better understanding of worker types and their specific intrinsic motivations is essential in driving the boundaries of current crowdsourcing quality. Kazai et al. [21] proposed an interesting classification of workers into several categories. In their work, a number of performance-based worker types, including e.g., spammers, sloppy and competent workers are described. We believe, that more general worker models which also encompass aspects such as worker motivation capability and interest in certain HIT types, etc. can be of significant benefit for the field. Very similar to the task of advertisement placement, a worker whose motivations we understand, can be targeted with better-suited precisely-tailored HIT types.

The common example of worker filtering by nationality illustrates the practical need for a better understanding of worker motivation. This practice is not only of dubious ethical value, it may additionally address symptoms rather than causes. The original objective of identifying and rejecting such workers that try to game the task and get paid without actually working is often hard to fulfil. Filtering by nationality is straightforward to achieve, but also (at best) only correlated with reliability. This bears the significant risk of artificially thinning the pool of available workers. This work (e.g., Tables 4 and 6), demonstrates that in an entirely unfiltered environment no significant national differences in quality, cheat rates, etc. could be found when focussing on the desired worker type. In this way, we retain a large work force but, by task design, discourage undesired worker types from taking up our work in the first place.

Looking out towards future changes to the game based on lessons learned in this work, we aim for including yet another incentive besides entertainment. The leaderboard concept of the current game tries to spark competition between players and has a moderate success at doing so. However, the workers do not know each other. In a reputation-aware environment, such as a social network, this effect can be expected to have a far greater impact. Having the ability to compare scores and to compete in a direct multi-player game with their friends will create much more compelling incentives for (a) performing well in each game, (b) continuing to play, (c) returning for subsequent matches and (d) recommending the game to their circle of friends. We believe that exploiting these aspects by integrating social reputation into crowdsourcing will create many interesting applications.

7. CONCLUSION

In this work, we demonstrated the benefits of a game-based approach for collecting relevance assessments, exploiting insights from the field of serious games for application in commercial large-scale crowdsourcing. After a description of key design criteria, we evaluated the proposed scheme following the setup of the TREC 2011 Crowdsourcing Track. We achieve confirm high result quality, matching the performance levels of the best TREC participants for the same task at a fraction of the invested cost, while attracting fewer cheaters. In a dedicated experiment, we showed the generalizability of the proposed game, that, without any changes, was applied for the task of image classification and clustering. In summary, we are convinced that alternative incentives besides the actual financial HIT reward can positively influence the outcome of crowdsourced data collection and annotation campaigns. As a tangible outcome of this work, a large-scale set of relevance judgements towards the T11Crowd subset of Clueweb was created and is available to the research community. Furthermore, the described game

can be accessed and deployed as an open source project[4]. Future work will further exploit the community aspect to increase the motivation for playing. This could be done by, e.g., introducing a multiplayer mode in which several players are in direct competition or by integrating the game into an identity and reputation-aware environment such as social networks or virtual worlds. More fundamentally, this work has demonstrated the benefit of addressing the specific preferences of entertainment-seeking workers. In the future, however, we should investigate formal worker models of worker motivation and capability to enable an optimal work distribution and representation for arbitrary worker types.

Acknowledgements

We would like to thank Jiyin He and the Fish4Knowledge project for providing us with the fish images and expert judgements.

8. REFERENCES

[1] Gambit - Payment Solutions for Virtual Currency. http://www.getgambit.com/, 2011.

[2] O. Alonso and R. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *ECIR*, 2011.

[3] O. Alonso and M. Lease. Crowdsourcing 101: putting the WSDM of crowds to work for you. In *WSDM*, 2011.

[4] O. Alonso, D.E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. In *ACM SIGIR Forum*, 2008.

[5] B. Baldwin and B. Carpenter. LingPipe. *Available from World Wide Web: http://alias-i. com/lingpipe*, 2003.

[6] C. Buckley and E.M. Voorhees. Retrieval evaluation with incomplete information. In *SIGIR 2004*. ACM.

[7] J. Callan, M. Hoy, C. Yoo, and L. Zhao. Clueweb09 data set. *http://boston.lti.cs.cmu.edu*, 2009.

[8] B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 268–275. ACM, 2006.

[9] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas. Million Query track 2009 overview. In *Proceedings of TREC*, volume 9, 2009.

[10] V.R. Carvalho, M. Lease, and E. Yilmaz. Crowdsourcing for search evaluation. In *ACM SIGIR Forum*, 2011.

[11] CrowdFlower. Crowdsourcing - Labor on demand. http://crowdflower.com/, 2012.

[12] M. Csikszentmihalyi. *Flow: The psychology of optimal experience*. Harper Perennial, 1991.

[13] C. Eickhoff and A. de Vries. How crowdsourcable is your task. In *WSDM Workshop on Crowdsourcing for Search and Data Mining (CSDM)*, pages 11–14, 2011.

[14] C. Eickhoff and A.P. de Vries. Increasing Cheat Robustness of Crowdsourcing Tasks. *Information Retrieval To appear*, 2012.

[15] C. Grady and M. Lease. Crowdsourcing document relevance assessment with Mechanical Turk. In *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[16] C. Harris. You're Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks. In *WSDM Workshop on Crowdsourcing for Search and Data Mining (CSDM)*, pages 15–18, 2011.

[17] M. Hirth, T. Hoßfeld, and P. Tran-Gia. Cheat-Detection Mechanisms for Crowdsourcing. *University of Würzburg, Tech. Rep*, 2010.

[18] P. Ipeirotis. Crowdsourcing using Mechanical Turk: quality management and scalability. In *Proceedings of the 8th International Workshop on Information Integration on the Web: in conjunction with WWW 2011*. ACM, 2011.

[19] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.

[20] G. Kazai. In search of quality in crowdsourcing for search engine evaluation. *ECIR*, 2011.

[21] G. Kazai, J. Kamps, and N. Milic-Frayling. Worker Types and Personality Traits in Crowdsourcing Relevance Labels. 2011.

[22] G. Kazai and N. Milic-Frayling. On the evaluation of the quality of relevance assessments collected through crowdsourcing. In *SIGIR 2009 Workshop on the Future of IR Evaluation*, 2009.

[23] C.H. Ko, J.Y. Yen, C.C. Chen, S.H. Chen, and C.F. Yen. Gender differences and related factors affecting online gaming addiction among taiwanese adolescents. *The Journal of nervous and mental disease*, 2005.

[24] J. Lampel and A. Bhalla. The role of status seeking in online communities: Giving the gift of experience. *Journal of Computer-Mediated Communication*, 2007.

[25] M. Lease and G. Kazai. Overview of the TREC 2011 Crowdsourcing Track (Conference Notebook). 2011.

[26] M. Lease and E. Yilmaz. Crowdsourcing for information retrieval. In *ACM SIGIR Forum*, number 2. ACM, 2012.

[27] H. Ma, R. Chandrasekar, C. Quirk, and A. Gupta. Improving search engines using human computation games. In *CIKM 2009*.

[28] C.C. Marshall and F.M. Shipman. The ownership and reuse of visual media. *JCDL*, 2011.

[29] J. McGonigal. *Reality is broken: Why games make us better and how they can change the world*. Penguin Pr, 2011.

[30] David Pennock. The Wisdom of the Probability Sports Crowd. http://blog.oddhead.com/2007/01/04/the-wisdom-of-the-probabilitysports-crowd/, 2007.

[31] C. Richards. Teach the world to twitch: An interview with Marc Prensky, CEO and founder Games2train. com. Futurelab, 2003.

[32] F. Scholer, A. Turpin, and M. Sanderson. Quantifying test collection quality based on the consistency of relevance judgements. In *SIGIR 2011*.

[33] 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 *EMNLP*, 2008.

[34] J. Surowiecki, M.P. Silverman, et al. The wisdom of crowds. *American Journal of Physics*, 2007.

[35] Amazon Mechanical Turk. Artificial Artificial Intelligence. http://mturk.com, 2012.

[36] J. Urbano, M. Marrero, D. Martín, J. Morato, K. Robles, and J. Lloréns. The University Carlos III of Madrid at TREC 2011 Crowdsourcing Track: Notebook Paper. 2011.

[37] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *SIGCHI*, 2004.

[38] L. von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 2008.

[39] L. von Ahn, S. Ginosar, M. Kedia, and M. Blum. Improving image search with phetch. In *ICASSP*, 2007.

[40] L. von Ahn, M. Kedia, and M. Blum. Verbosity: a game for collecting common-sense facts. In *SIGCHI 2006*.

[41] L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *SIGCHI 2006*.

[42] E. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*, 2002.

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

[44] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *NIPS*, 2010.

[4] http://sourceforge.net/projects/geann/