RMIT at the Crouwdsourcing Track of TREC 2011

Matthias Petri
RMIT University
School of CS&IT
matthias.petri@rmit.edu.au

Mark Sanderson
RMIT University
School of CS&IT
mark.sanderson@rmit.edu.au

Falk Scholer
RMIT University
School of CS&IT
falk.scholer@rmit.edu.au

***Abstract*—In this paper we describe our submission to the crowdsourcing track of TREC 2011. We first describe our crowdsourcing environment. Next we evaluate our approach and discuss our results. We conclude with a discussion of problems encountered during our participation.**

I. Introduction

Crowdsourcing has become a useful tool that is used by the research community to parallelize human interaction tasks. Many experiments performed by university affiliated IR researchers tend to only include very few participants, usually students, due to the limit amount of funds available in comparison to companies like Google. Crowdsourcing experiments, on the other hand, can be performed on a larger scale, with a very diverse group of participants and at a reasonable cost.

The purpose of our participation in the 2011 crowdsourcing track is to evaluate crowdsourcing as an alternative to traditional user-focused IR experiments. In this paper we first describe our experimental setup under the constraints of the trec guidelines. Next we evaluate our submitted results and conclude with a discussion of problems encountered during our participation.

II. Experimental Design

The TREC 2011 Crowdsourcing Track consisted of two tasks: an assessment task, and a consensus task. RMIT University participated in the first of these tasks, the aim of which was to evaluate the effectiveness of crowdsourding to collect relevance judgements. The assessments were made on documents from a subset of the ClueWeb test collection, a 25TB crawl of the World Wide Web in 2009. To avoid the possibility of harmful code, documents were rendered as image files.

In our experiment we used CrowdFlower (CF) to create and manage our assessment tasks. We used the CF internal markup language to define our interface. The basic design of a single human intelligence task (HIT) is shown in Figure 1. Query terms, description and narrative were extracted from the TREC topic descriptions and displayed above each image. When possible, the "shortened" version of the website image was used to decrease loading times. Instead of showing the complete image directly we used the crowdflower webservers to provide a clickable, cached thumbnail of size 800x400. Note that CF's validator functionality was used to ensure all workers select either relevant or not relevant in the voting widget.

The TREC guidelines require every worker to judge all 5 documents of a given topic–document set. However, we defined one HIT to only consist of one image relevance judgement. To ensure that individual workers perform all image relevance judgements in a topic set we specified, in the CF job interface, that 5 HITs per page should be displayed at the same time.

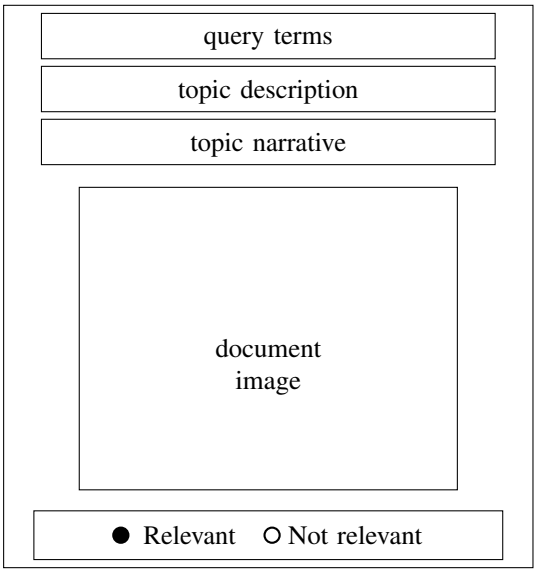

Figure 1. Layout of one HIT unit consisting of query terms, description, narrative, image thumbnail and relevance voting widget.

A. Quality Assurance

We used the CF internal gold data facilities to control the quality of our workers. We manually created 23 gold standard image/topic sets containing 5 images each which were chosen to be clearly identifiable as either relevant or non-relevant. For a given topic, we selected documents which were either judged relevant to a different topic or could easily be identified as not relevant or relevant by simple visible features.

A worker is periodically presented with a gold standard set of 5 image relevance judgement tasks for which we predefined the correct answer. The worker gets notified if he incorrectly answers any of the test tasks. If an individual worker fails to answer several of these gold standard questions we no longer present him with additional HITs.

Crowdflower categorizes workers into two categories: trusted and and "not trusted". Trusted workers are shown

gold standard answers sets less frequently than "not trusted" workers.

B. Pricing

We gathered 2 judgements per document set. In total we gathered 866 judgements over 456 topic - document sets at a cost of $133.43.

III. Evaluation

The task was "ordered" at 9pm AEST and finished at roughly 11pm on the same day. Overall 1376 trusted judgements and 424 untrusted judgements were gathered. Judgements were collected from 44 workers.

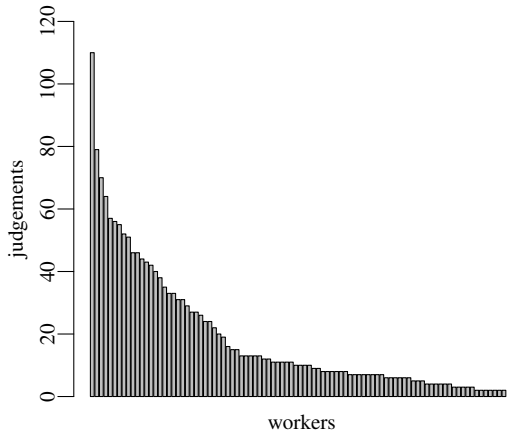

Figure 2. Worker judgement distribution

The judgement/worker distribution can be seen in Figure 2. Note that 15 workers performed 50% of all tasks. Table I shows the country worker distribution.

Country	% of judgements
India	77%
Serbia	5%
Philippines	4%
Japan	3%
Thailand	2%
USA	2%
Other	6%

Table I
JUDGEMENT COUNTRY DISTRIBUTION.

CF only reports the time taken to complete 1 HIT (5 document judgements) in minutes. We can therefore only give rough estimates on the time it took workers to complete on set of documents: 28% of the sets were completed in less than one minute. 50% were completed in $1-2$ minutes. 12% were completed in $2-3$ minutes. The other 6% of all judgements were completed in $3-22$ minutes.

A. Worker quality

The median average worker trust calculated by CF was 0.9. Only 12% of our gold standard test sets were answered incorrectly. We first use consensus based evaluation to judge worker quality. Overall our worker quality is similar

Team	Accuracy	Recall	Precision	Specificity
BUPT-WILDCAT	75.7%	83.8%	76.3%	64.2%
GeAnn	65.0%	76.9%	66.8%	45.4%
MSRC	77.0%	70.7%	86.5%	83.3%
RMIT	76.4%	76.6%	80.2%	75.4%
TUD_DMIR	65.6%	63.1%	74.4%	67.8%
uc3m	72.0%	71.8%	77.6%	72.2%
uogTr	78.2%	**86.6**%	80.4%	64.0%
UwaterlooMDS	**81.9**%	73.7%	**90.5**%	**91.2**%
Mean	73.97%	75.4%	79.08%	70.43%

Table II
TASK 1 CONSENSUS RESULTS.

Team	Accuracy	Recall	Precision	Specificity
BUPT-WILDCAT	**91.2**%	**97.6**%	**91.9**%	**73.4**%
GeAnn	62.3%	74.8%	72.4%	26.5%
MSRC	65.0%	64.3%	79.1%	62.0%
RMIT	58.2%	63.6%	72.8%	47.2%
TUD_DMIR	62.0%	64.6%	77.3%	53.6%
uc3m	72.3%	72.9%	85.3%	71.9%
uogTr	61.7%	73.8%	67.7%	33.0%
Mean	67.5%	73.0%	78.07%	52.51%

Table III
TASK 1 GOLD RESULTS.

to that of other groups using crowdsourcing. We however always perform better than the mean consensus score over all submitted runs. The runs using crowdsourcing tend to perform worse than a submitted single worker run (UWaterlooMDS).

Next we evaluate worker quality by comparing to existing TREC gold data. Team BUPT-WILDCAT outperforms the other teams on all metrics. Our run performs worse than the mean of all submitted runs. There is a large discrepancy between the performance of our run in the consensus evaluation and in the gold standard evaluation.

IV. Conclusion

We performed binary relevance judgements of images using CrowdFlower. We used gold standard data to control worker quality. However, the gathered judgements are only 76.4% accurate even though our measured worker quality through CF is very high. We conjecture that our gold standard data was not "good enough" to filter *all* non-trustworthy workers. Overall, judgements obtained using crowdsourcing are of similar quality to that of a single high quality worker.