Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment

Omar Alonso
A9.com
Palo Alto, CA (USA)
oralonso@gmail.com

Stefano Mizzaro
Dept. of Mathematics and Computer Science
University of Udine
Udine (Italy)
mizzaro@dimi.uniud.it

ABSTRACT

Recently, Amazon Mechanical Turk has gained a lot of attention as a tool for conducting different kinds of relevance evaluations. In this paper we show a series of experiments on TREC data, evaluate the outcome, and discuss the results. Our position, supported by these preliminary experimental results, is that crowdsourcing is a viable alternative for relevance assessment.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and software — performance evaluation

General Terms

Measurement, performance, experimentation

Keywords

IR evaluation, relevance, relevance assessment, user study

1. INTRODUCTION AND MOTIVATIONS

One issue in current TREC-like test collection initiatives is the cost related to relevance assessment: assessing requires resources (that cost time and even money) and does not scale up. Indeed, in recent years, there has been some trend on trying to save assessment resources: there is a vast body of literature on reducing the number of documents pooled and/or judged, and, more recently, on reducing the number of assessed topics [4] as well. Also, test collections are sometimes built in-house [3], and assessment effort is obviously a problematic issue when building your own test collection.

Stated briefly, our research question is: "Can we get rid of TREC assessors?" Our position is that crowdsourcing is a reliable alternative to "classical" assessors: in this paper we propose to use the Mechanical Turk crowdsourcing platform for relevance assessing; we also support this approach by some experimental data.

2. RELATED WORK

Amazon Mechanical Turk (MTurk, `www.mturk.com`) is a marketplace for work that requires human intelligence. The individual or organization who has work to be performed is known as the requester. A person who wants to sign up to perform work is described in the system as a turker. The unit of work to be performed is called a HIT (Human Intelligence Task). Each HIT has an associated payment and an allotted completion time. It is possible to control the quality of the work by using qualification tests. MTurk has already been used in some relevance related research [1,2,5], with good success.

E1	Graded relevance on a 4 point scale (3 = excellent, 2 = good, 1 = fair, 0 = not relevant) following closely TREC-7 guidelines. We summarized the main points from the TREC assessment guidelines as starting point.
E2	Graded relevance with modified instructions. Changes on the instructions, use more layman English (not so expert). We also included an input form in the task so turkers can provide feedback.
E3	Graded relevance with modified instructions II. Modified instructions using colors and examples of relevant content. Also included more documents in the test.
E4	Binary relevance without qualification test. Maintained same instructions but changed the answers to binary (1 = relevant and 0 = not relevant). Modified the feedback input to an optional entry for justifying answers. Passing grade was 80% of correct answers.
E5	Binary relevance with qualification test. Same as previous experiment but with a lower passing grade for the qualification test to 60%.

Table 1: The five experiments

Therefore, our research question can be framed as: "Is it possible to replace TREC-like relevance assessors with Mechanical turkers?". We think the answer is "Yes — at least to some extent"; we report in the next sections some preliminary experimental results that support our position.

3. EXPERIMENTS

We used the TREC topic about space program (number 011), in the domain of science and technology. We selected a subset of 29 FBIS documents (the first 14 not relevant, and the first 15 relevant). Each turker was given some instructions, a description of the topic, and one document, and he was asked to judge the relevance of the document to the topic. We decided to have each topic/document pair judged by 10 turkers, thus obtaining 290 judgments in total (when the task was 100% complete).

We performed 5 experiments, as shown in Table 1. We refined the experiments and methodology in each experiment run in an iterative way. By looking at the results data, we

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
.

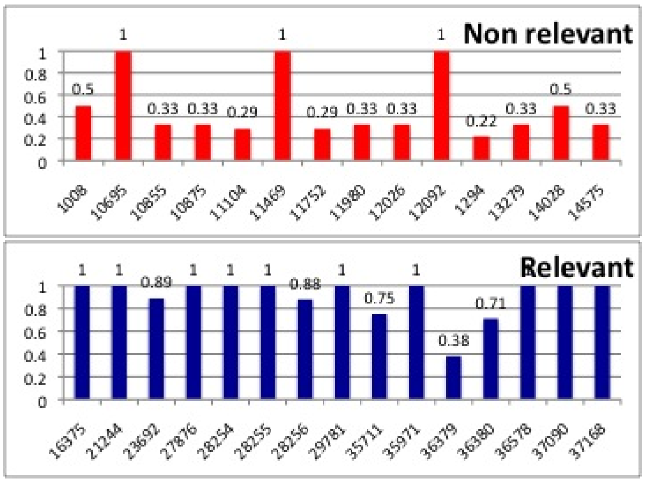

Figure 1: Turkers average relevance assessments

manually inspected the answers, adjusted the methodology accordingly, and tested again. This was done over several weeks as the completion for each experiment was set to 10 days. For each experiment we paid 0.02 cents per task.

The task design in MTurk can be framed as a user interface problem, so in every iteration we tweaked the language, instructions, and presentation. As the results looked closer to our initial hypothesis, we decided to use binary evaluation with qualification test. For this particular topic (space program), we felt that binary evaluation was more suitable given the content of the collection.

We measured the agreement between the turkers and TREC assessors as presented in Figure 1 (that shows the FBIS3 documents on the X axis and the average turkers score on the Y axis, with relevant = 1 and not relevant = 0). For the relevant documents the average across all turkers was 0.91 (relevant expert assessment was 1). In the case of not relevant document, the average was 0.49 (not relevant expert assessment was 0). There are 4 exception where turkers disagree with the experts, for documents: 10695, 11469, 12092, and 36379. We manually inspected the documents and concluded that, in three out of four cases, turkers were correct and TREC assessors were wrong: document FBIS3-10695 seems definitely relevant; 11469 is probably not relevant, but partially relevant; 12092 sounds relevant; and 36379 is not relevant.

Of all the assignments in E5, 40% contain turker's justifications for answers. This feedback field was not mandatory in the experiment. In most of the cases turkers provided a very good explanation. For example, concerning not relevant documents:

- This report is about the Russian economy, not the space program.
- The "MIR" in the article refers to a political group, not the Russian space station.
- This article is about Kashmir, not the space program.

And concerning relevant ones:

- This is about Japan's space program and even refers to a launch.
- On the Russian space program, not US, but comments about American interest in the program.
- The article is relevant, but it seems a non-native English speaker wrote it. For instance the article says the space shuttle will lift off from the "cosmodrome". NASA doesn't call the launch pad a "cosmodrome."

4. DISCUSSION AND OUTCOMES

As we can see from the data analysis, turkers not only are accurate in assessing relevance but in some cases were more precise than the original experts. Also, turkers tend to agree slightly more with the experts when the document is relevant, and less when it is not relevant.

It is important to design the experiments carefully. Mapping TREC assessment instructions [6] to MTurk is not trivial. The TREC-7 guidelines is a 4-page document that has to be summarized in a few sentences for reading online, since the turker sees a screen with instructions and task to be completed. It is important to be concise, precise, and clear about how to evaluate the relevance of a document. The usage of some basic usability design considerations for presentation is also important.

In our experience, all experiments without qualification tests were completed in less than 48 hours. Once qualification test was involved, the completion rate per turker was much higher. The number of turkers required to assess per document can have an impact on the duration.

5. CONCLUSIONS

Crowdsourcing-based relevance evaluation using MTurk is a feasible alternative to perform relevance evaluations. Using TREC data, we have demonstrated that the quality of the raters is as good as the experts. Our experience shows that it is extremely important to carefully design the experiment and collect feedback from turkers. Taking a TREC-like experiment and run it as is, would probably fail.

In the future, we plan to seek confirmation of these findings on more TREC topics, and also to study related issues like the effect of topics/documents of different kinds, the number of turkers needed to get reliable results, the importance of the qualification test etc.

6. REFERENCES

[1] O. Alonso and S. Mizzaro. Relevance Criteria for E-Commerce: A Crowdsourcing-based Experimental Analysis. In *Proceedings of SIGIR'09*, 2009. In press.

[2] O. Alonso, D. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[3] J. Callan, J. Allan, C. L. A. Clarke, S. Dumais, D. A. Evans, M. Sanderson, and C. Zhai. Meeting of the MINDS: An Information Retrieval Research Agenda. *SIGIR Forum*, 41(2):25–34, 2007.

[4] J. Guiver, S. Mizzaro, and S. Robertson. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM TOIS*, 2009. In press.

[5] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *CHI '08: Proceeding of the 26th SIGCHI*, pages 453–456, 2008.

[6] E. Voorhees. Personal communication.