ERICA: Expert guidance in validating crowd answers

Nguyen Quoc Viet Hung\textsuperscript{*}, Duong Chi Thang\textsuperscript{†}, Matthias Weidlich\textsuperscript{*}, Karl Aberer\textsuperscript{†}

\textsuperscript{*}École Polytechnique Fédérale de Lausanne, \textsuperscript{†}Imperial College London

ABSTRACT
Crowdsourcing became an essential tool for a broad range of Web applications. Yet, the wide-ranging levels of expertise of crowd workers as well as the presence of faulty workers call for quality control of the crowdsourcing result. To this end, many crowdsourcing platforms feature a post-processing phase, in which crowd answers are validated by experts. This approach incurs high costs though, since expert input is a scarce resource. To support the expert in the validation process, we present a tool for Expert guidance In validating Crowd Answers (ERICA). It allows us to guide the expert’s work by collecting input on the most problematic cases, thereby achieving a set of high quality answers even if the expert does not validate the complete answer set. The tool also supports the task requester in selecting the most cost-efficient allocation of the budget between the expert and the crowd.

1. INTRODUCTION
Crowdsourcing has been established as an efficient and scalable approach to overcome problems that are computationally expensive or unsolvable for machines, but rather trivial for humans \cite{6}. In essence, crowdsourcing enables users to post tasks in the form of questions, which are then answered by crowd workers for financial rewards. Crowd workers, however, have different backgrounds and wide-ranging levels of expertise and motivation \cite{3}, so that the obtained answers are not necessarily correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answer are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposedly correct. Even aggregation of answers from multiple workers to a single question cannot guarantee result correctness. Platforms for crowdsourcing, such as Amazon Mechanical Turk (AMT), aim at including the answer trustworthiness by including a validation phase. That is, the obtained crowd answers are checked against expert answers that are supposed...
partitions the expert feedback into test and training sets. Then, the training set is used to compute the aggregated answers for the test set and the similarity between two aggregated answers sets (one created using only the training set, one using the whole feedback set) is measured. Repeating this procedure \(k\) times, ERICA estimates the quality of the aggregated answers.

**Optimizing cost.** ERICA helps to find the optimal distribution of a pre-defined budget between an expert and the crowd based on characteristics of the crowd. Since these characteristics are not known beforehand, it samples the crowd with a few test questions and simulates the crowd to decide on the best budget distribution [5].

**Detecting expert mistakes.** ERICA also uses \(k\)-fold cross validation to detect expert mistakes. Here, \(k\) is fixed to 1, though, in order to detect incorrect feedback. Then, the aggregated answer for the question in the test set is compared with the feedback by the expert. If they are different, the expert is encouraged to revisit the respective question and reconsider the answer.

**Scalability**

To achieve an efficient implementation for large-scale data, ERICA employs two techniques: *parallelization* and *partitioning*. First, since the computations of the information gain and the spammer score are independent for different questions, they are executed in parallel concurrently for all questions. Second, common answer score are independent for different questions, they are executed in parallel. This leads to speedup of factor \(k\).

**Technicality**

ERICA is implemented as a web service. The back-end is written in Python and Java. The front-end is an HTML5 website communicating with the server using Javascript and jQuery.

### 3. DEMONSTRATION

To demonstrate the application of ERICA, we first present its user interface before we turn to an example scenario.

**User interface.** ERICA offers a rich user interface (see Figure 1), which consists of 2 main components:

- **Answer matrix:** contains all the questions and crowd answers. Each row is a question, each column is a worker and a cell denotes the respective answer. The workers and the questions are color-coded to reflect their reliability and difficulty, respectively. The aggregated answers, considered to be correct according to the crowd answers and the expert feedback, are shown in a separate column.
- **Expert feedback view:** contains three panels that present additional information during the validation. Panel A shows the content of the currently-selected question that requires expert input. Various statistics are provided in Panel B for the expert’s consideration. Further, the history of expert feedback is visualized in Panel C.

**Demonstration scenario.** We use the example from Figure 1 to demonstrate how to use ERICA to analyze and validate the answer matrix (a screencast of the demonstration is publicly available[1]).

Before posting a task to a crowdsourcing platform, a user may compute the budget allocation between an expert and the crowd. First, a user posts a set of test questions to study the characteristics of the crowd. Based on the answers, ERICA will simulate the crowd and compute the best budget allocation. Then, the user decides how many questions are posted to the crowd.

After all answers have been collected for the crowdsourcing task, the expert starts the validation. First, ERICA will select the question, for which the feedback is most beneficial. The question and its possible answers are shown in Panel A. Once the expert has submitted their answer, ERICA will recompute the aggregated answers and the reliability of the workers. After several feedback iterations, the expert can remove the answers of a worker, if the estimated reliability of the worker is too low. ERICA will measure the effect of this removal and recompute the result accordingly. The expert can stop the validation process when the estimated precision shown in Panel B is high or the aggregated answers and the feedback are similar for a consecutive number of times, as shown in Panel C.

### 4. SUMMARY

This paper presented ERICA: a tool for expert guidance in validating crowd answers. Unlike traditional tools for crowdsourcing, ERICA is the first to streamline the validation process by an expert for crowd answers. The expert feedback provides a reliable means to detect and remove faulty workers, which can improve the quality of the result significantly. ERICA helps crowdsourcing users by computing the best budget allocation between the expert and the crowd. As such, ERICA serves not only as a tool to reduce the validation effort, but also as a decision-support system to achieve a reduction of validation cost and an improvement of result quality.

### References


https://code.google.com/p/crowdvalidator/