Towards Quality Assessment of Crowdworker Output Based on Behavioral Data

Shigeaki Yuasa*, Takumi Nakai†, Takanori Maruichi‡, Manuel Landsmann§, Koichi Kise¶
Dept. of Computer Science and Intelligent Systems, Osaka Prefecture University, Japan
{*yuasa, †tnakai, ‡maruichi}@m.cs.osakafu-u.ac.jp, §m_landsman09@gmx.de, ¶kise@cs.osakafu-u.ac.jp

Masaki Matsubara‖, Atsuyuki Morishima**
Faculty of Library, Information and Media Science, University of Tsukuba, Japan
{‖masaki, **mori}@slis.tsukuba.ac.jp

***Abstract*—In this paper, we show preliminary results on the quality assessment of crowdworker output based on the movements of the mouse and the eyes while the task is performed. We assume that the mouse and the eyes stop longer if the quality is lower due to the lack of knowledge, or confidence, etc. Because the mouse- and eye-stopping duration follows log-normal distribution, we estimate its parameters (mean and standard deviation) to evaluate the quality. Results of preliminary experiments with 10 participants show that the parameters of correct outputs are different from those of incorrect ones. As compared to the task duration, which is often used as a feature for assessment, we have found that the mouse- and the eye-stopping duration is advantageous and complementary for the assessment.**

***Index Terms*—quality assessment, crowdsourcing, eye tracking, mouse movement, log-normal**

I. Introduction

Quality control has been recognized as an essential task in the field of crowdsourcing. A fundamental step of quality control is the assessment of quality, which allows us to take further actions for improvement of quality. In this paper, we report some preliminary results of the joint project on the quality assessment by two universities under the umbrella of JST CREST projects: Experiential Supplements for sensing and actuating human behavoirs [1], [2], and CyborgCrowd to build a crowdsourcing platform for both AI and humans.

Although many researchers have proposed quality assessment for eliminating spam workers, a limited number of efforts have been made for evaluating serious (non-spam) workers. When we launch research on this topic, the first question we need to consider could be about how to obtain the information for quality assessment. Daniel *et al.* have classified methods of quality assessment into the following three categories: individual, group, and computation-based [3]. Methods of the category “individual” indicate the assessment by an individual. The category “group” is about the assessment by a group of people. In contrast to the above methods, “computation-based” is to assess the quality without the involvement of humans. We focus here on the “computation-based” assessment.

The quality has been assessed on various targets such as workers, tasks, outputs, and environments. In this paper, we are concerned with the assessment of outputs: whether the result of the task performance is correct or not. Methods of assessing the quality of outputs can be classified into two broad categories: methods based on worker outputs, and those based on worker behaviors [4]. We are working on the assessment based on worker behaviors since they include rich information relevant to the accuracy: whether the worker immediately selects the answers without hesitation, or not. In particular, we employ the information obtained from mouse movement and eye movement to differentiate the workers’ behaviors with correct and incorrect outputs.

Our method is based on the assumption that the mouse and the eyes stop longer during the performance of a task if the worker is with less knowledge or confidence to the output resulting in an incorrect answer. The contribution of this paper is as follows:

- As compared to the task duration, which is often used as a feature of assessment, the cumulative duration of mouse stops, and the cumulative duration of eye stops (fixations) are informative and seem to be more robust in some cases.
- The stopping duration of the mouse and the eyes are complementary. The eye-stopping duration is more informative in some cases, and vice versa.

II. Related Work

Fingerprinting is a method that can be used for assessing worker quality by means of their behavior instead of their output. One specific case in which fingerprinting is used in crowdsourcing is to identify attackers among product reviewers on e-commerce websites [5]. For this purpose, the logarithm of the length of the written reviews and some account details of the reviewer are used. Further, reviewers are clustered to find similar reviewers and possible collaborating attackers.

Rzeszotarski *et al.* use a more general approach with fingerprinting [4]. They showed that features extracted from the mouse and keyboard input could be used in various crowdsourcing tasks to differentiate between high and low quality workers. For example, long periods of time with no input might result from distractions, while too quick executions might indicate a lack of serious effort. The quality was depending on the task quantified either over correct answers or over the amount and meaningfulness of labels written for images. In follow-up work, they offer a visualization tool that combines the aforementioned fingerprinting with worker output [6]. It allows us to manually assess workers and output from complex or creative crowdsourcing tasks. The definition of the quality criteria is up to the requester.

For simple human intelligence tasks (HIT), there is an additional challenge that the workers' interaction during a task is too short to get enough data about their behavior. Suzuki *et al.* evade this problem by adding extra operations into a task for the worker [7]. However, such measures can make the work more tedious and slow down the workers.

Another solution could be to use additional means of observing the workers, for example, with eye trackers. Hence, we are working on combining features from mouse input and eye tracking to offer a new and accurate way of assessing workers' behavior.

III. Behavior Analysis

A. Data Recording

We employed 10 workers (university and graduate school students, male: 10, age: 20's), then asked them to solve simple HITs tasks. Workers who completed the tasks received 1,000 JPY. We recorded the mouse movement and the eye movement behavior with a laser mouse and a stationary eye tracker (Tobii nano pro, 60Hz). The details of the task are as follows.

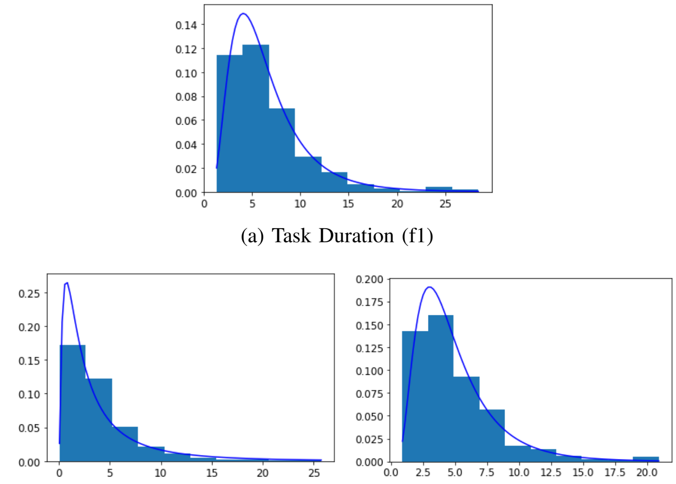
Two bibliographic descriptions of books consisting of the title, volume, author, publisher, and year of publication were displayed. Some information may be missing or contains small errors. The worker's task was to decide whether those two book descriptions were about the same book or two different ones. After clicking on one of two buttons representing "same" and "different," the next two descriptions were shown. Each worker was asked to solve 50 tasks. All workers did the same tasks in the same order.

B. Features

As features of analyzing worker behaviors, we employ the following three features about duration: f1: the task duration which is the elapsed time from the start to the end of a task, f2: cumulative duration of mouse stops during the task, and f3: cumulative duration of eye stops (fixations) during the task. We expect that these features reflect workers' uncertainty and perplexity [1], [2]. The details of f2 and f3 are as follows.

For f2, x and y coordinates are recorded approximately in 2Hz. When the moving distance is within 1 pixel, we regarded it as a "mouse stop". We sum up the duration of mouse stops, then set it as the feature f2.

As for f3, we record x and y coordinates of gaze points. We classify these data into two events: fixations and saccades by using an algorithm proposed by Buscher *et al.* [8]. We sum up the duration of all fixations during a task and set it as the feature f3.



(a) Task Duration (f1)

(b) Mouse (f2)

(c) Eye (f3)

Fig. 1: Distributions of features. The horizontal and the vertical axes indicate duration [s] and relative frequencies, respectively.

C. Log-normal Model and Behavior Representation

It is known that f1 follows the log-normal distribution [9], which is defined as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\},$$

where μ and σ are the mean and the standard deviation of the logarithm of feature x. This distribution often appears in the natural world [10].

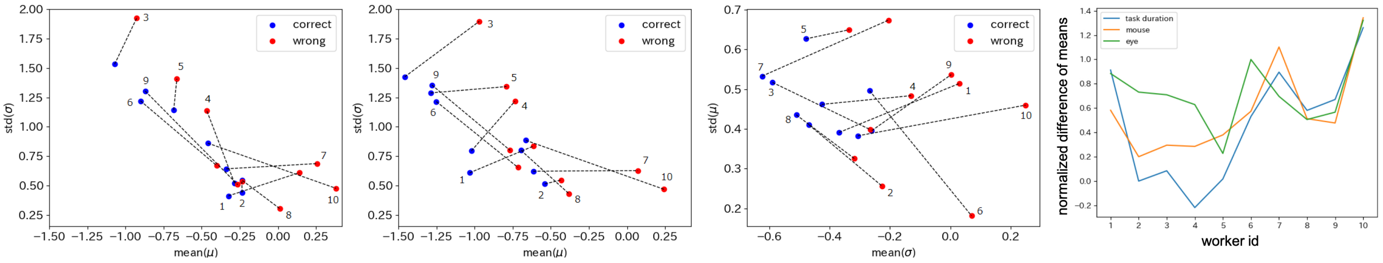
We investigated whether f2 and f3 are also following the log-normal distribution, and found that they do as shown in Figs. 1b and 1c. It is also confirmed that the task duration follows log-normal distribution in our dataset (see Fig. 1a).

We consider that the above fact can be used to highlight the behavior by calculating the parameters μ and σ, assuming that different values of parameters can be obtained from the behaviors with correct and incorrect outputs.

IV. Experimental Results and Discussion

We asked all workers to complete 50 tasks. Then we put the labels of correct or incorrect for each output by using the ground truth. The ratio of correct and incorrect outputs is approximately 4:1. The three features were calculated for the correct and incorrect outputs of each worker. First, the values of each feature for either the correct or incorrect outputs were summed and divided by the mean of all data from a worker. Then these normalized values for all tasks were employed to produce the representation by the log-normal model. Thus for each worker, the representations of correct and incorrect outputs by μ and σ were obtained.

The result is shown in Fig. 2. Blue and red points show the correct and incorrect outputs of a worker, respectively. The outputs of the same worker are connected with a dotted line. The numbers in the figures show worker IDs.The quality



(a) Task Duration (f1) (b) Mouse (f2) (c) Eye (f3) (d) Normalized difference

Fig. 2: Representation of each feature. In (a)–(c), the horizontal and the vertical axes represent μ and σ of the log-normal, respectively. The numbers 1–10 in the figures indicate workers. (d) represents the difference of means of correct and incorrect outputs divided by the standard deviation.

can be assessed if means of correct and incorrect answers are different enough in terms of their standard deviations.

As you can see in the figures, most of the correct outputs are separated from the incorrect ones in all features. However, for some workers such as workers 2, 3, and 5, we can only find a small difference in the feature task duration (f1). For worker 2, there is almost no difference. For other workers 3 and 5, there exist gaps, but only in the standard deviation; small gaps are found in the mean of these data. This indicates the difficulty to classify outputs by only using f1.

On the other hand, f2 and f3 allow us to differentiate outputs for workers 3 and 5 with larger gaps in mean (μ). For worker 5, a more apparent separation is achieved by f2. For worker 2, f3 enables us to separate outputs, though f2 does not.

Fig. 2d shows the normalized difference of means, i.e., the difference of means of correct and incorrect outputs divided by the standard deviation of data for each worker. The larger the normalized difference is, the clearer the separation is. From this figure, f2 and f3 features are more advantageous in separation than f1.

To sum up, the mouse (f2) and the eye movement (f3) features are more effective in some workers than the task duration feature. In addition, these two features are found to be complementary.

V. Conclusion

We have presented the preliminary results of quality assessment based on the movement of mouse and eyes. In our method, we have focused on the cumulative duration of the mouse and the eye stops based on the assumption that it reflects time to consider, dither, and making a decision, which could be relevant to the correctness of the answer. The log-normal model has been employed to highlight the difference in duration between correct and incorrect answers. The experimental results show that, as compared to the task duration, which is often used for characterizing worker's behavior, the above two kinds of duration are more informative and complementary.

Future work includes further analysis of distribution to make it sure that it can be fruitful for predicting the accuracy of worker output, as well as to build a system to control the quality of crowdsourcing based on it.

Acknowledgment

This work was supported in part by JST CREST JPMJCR16E1, and JPMJCR16E3 including AIP challenge, Japan.

References

[1] K. Yamada, K. Kise, and O. Augereau, "Estimation of confidence based on eye gaze: an application to multiple-choice questions," in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM, 2017, pp. 217–220.

[2] T. Maruichi, K. Kise, O. Augereau, and M. Iwata, "Keystrokes tell you how confident you are: An application to vocabulary acquisition," in *Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2018 ACM International Symposium on Wearable Computers*. ACM, 2018, pp. 154–157.

[3] F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, and M. Allahbakhsh, "Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions," *ACM Computing Surveys (CSUR)*, vol. 51, no. 1, p. 7, 2018.

[4] J. M. Rzeszotarski and A. Kittur, "Instrumenting the crowd," in *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*. ACM Press, 2011.

[5] A. Fayazi, K. Lee, J. Caverlee, and A. Squicciarini, "Uncovering crowdsourced manipulation of online reviews," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15*. ACM Press, 2015.

[6] J. Rzeszotarski and A. Kittur, "Crowdscape: interactively visualizing user behavior and output," in *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 2012, pp. 55–62.

[7] Y. Suzuki, Y. Matsuda, and S. Nakamura, "Additional operations of simple HITs on microtask crowdsourcing for worker quality prediction," *Journal of Information Processing*, vol. 27, no. 0, pp. 51–60, 2019.

[8] G. Buscher, A. Dengel, and L. van Elst, "Eye movements as implicit relevance feedback," in *Proceedings of the 2008 CHI Conference on Human Factors in Computing Systems: Extended Abstracts*. ACM, 2008, pp. 2991–2996.

[9] P. Kucherbaev, F. Daniel, S. Tranquillini, and M. Marchese, "Relauncher: crowdsourcing micro-tasks runtime controller," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 2016, pp. 1609–1614.

[10] E. L. Crow and K. Shimizu, *Lognormal distributions*. Marcel Dekker New York, 1987.