Future Generation Computer Systems 87 (2018) 159–170

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

# Agreement-based credibility assessment and task replication in human computation systems

Lesandro Ponciano [a,b,*], Francisco Brasileiro [a]

[a] Federal University of Campina Grande, Bairro Universitário, Campina Grande, Paraíba, CEP 58429-900, Brazil
[b] Pontifical Catholic University of Minas Gerais, Bairro Coração Eucarístico, Belo Horizonte, Minas Gerais, CEP 30535-901, Brazil

## HIGHLIGHTS

- Human computation is analysed from the perspective of task replication.
- An adaptive credibility-based task replication algorithm is proposed.
- Four metrics of credibility of participants are proposed.
- Adaptive algorithm reaches accuracy similar to non-adaptive one, using fewer replicas
- Difficulty of tasks affects participants' credibility and algorithm performance.

## ARTICLE INFO

*Article history:*
Received 16 March 2017
Received in revised form 18 February 2018
Accepted 12 May 2018
Available online 22 May 2018

*Keywords:*
Human computation
Credibility
Task replication
Inter-rater agreement

## ABSTRACT

Human computation systems harness the cognitive power of a crowd of humans to solve computational tasks for which there are so far no satisfactory fully automated solutions. To obtain quality in the results, the system usually puts into practice a task replication strategy, i.e. the same task is executed multiple times by different humans. In this study we investigate how to improve task replication considering information about the credibility score of participants. We focus on how to automatically measure the credibility of participants while they execute tasks in the system, and how such credibility assessment can be used to define, at execution time, the suitable degree of replication for each task. Based on a conceptual framework, we propose (i) four alternative metrics to measure the credibility of participants according to the degree of agreement among them; and (ii) an adaptive credibility-based task replication algorithm that defines, at execution time, the degree of replication for each task. We evaluate the proposed algorithm in a diversity of configurations using data of thousands of tasks and hundreds of participants collected from two real human computation projects. Results show that the algorithm is effective in optimising the degree of replication, without compromising the accuracy of the obtained answers. In doing so, it improves the ability of the system to properly use the cognitive power provided by participants.



## 1. Introduction

Human computation is an emerging computing approach that draws upon human cognitive abilities to solve computational tasks for which there are still no satisfactory fully automated solutions [1–3]. Systems based on human computation are distributed systems that harness the cognitive power of a crowd of humans connected to the Internet to execute relatively simple tasks, whose solutions, once grouped, solve a problem that distributed systems equipped with only machines cannot solve satisfactorily. Such type of system has been proved to be effective in solving tasks that rely on human cognition such as detecting information in images [4], and processing natural language content [5], as well as more subjective tasks related to human's opinions and preferences [6].

There are currently two main types of human computation systems: *online labour markets* and *crowdsourced citizen science projects*. Online labour markets gather a crowd of humans willing to perform tasks in exchange for a relatively low financial incentive [7–9] — e.g. Amazon Mechanical Turk (mturk.com) and CrowdFlower (crowdflower.com). Crowdsourced citizen science projects, in turn, consist in a partnership between scientists and a crowd of humans willing to contribute to a scientific research, without receiving any financial incentive [10–12]. People acting in a citizen science project may contribute in a number of activities, which include performing human computation tasks. Examples of

* Corresponding author at: Pontifical Catholic University of Minas Gerais, Bairro Coração Eucarístico, Belo Horizonte, Minas Gerais, CEP 30535-901, Brazil.
*E-mail addresses:* lesandrop@pucminas.br (L. Ponciano), fubica@dsc.ufcg.edu.br (F. Brasileiro).

https://doi.org/10.1016/j.future.2018.05.028
0167-739X/

citizen science projects based on human computation are Stardust@home (stardustathome.ssl.berkeley.edu), in which people search for tiny interstellar dust impacts in images, and Galaxy Zoo (galaxyzoo.org), in which people perform morphological classification of galaxies from images.

To ensure quality in the execution of tasks, human computation systems usually require that the same task is executed multiple times by different humans; then, different aggregation mechanisms, which leverage the diversity and redundancy of multiple answers, can be employed to generate a more reliable answer to the task. In many systems, a task replication strategy is used as a way to obtain redundancy of answers in order to identify consensus in the set of answers or to tolerate faults that may cause some humans to generate wrong answers [11,13,3,14]. The *degree of replication* is the number of different humans who are performing each task. It is usually defined by the users at the moment of submitting a group of related tasks, all of them having the same degree of replication. Defining the suitable degree of replication for a task is a challenging process because it generates a trade-off between quality and cost. If the degree of replication is overestimated, an excessive amount of humans is used and, therefore, there is an increase in the cost of executing all tasks, perceived either financially or as a waste of resources that could have been allocated to do something else. On the other hand, if the degree of replication is underestimated, the desired quality in each answer is not achieved. Because tasks may differ among themselves in several ways, including its difficulty, it is expected that the ideal degree of replication can be different from one task to another, even when the description of the tasks are very similar. Given that users typically submit groups of hundreds or thousands of tasks, it is prohibitive to define manually a replication degree for each task addressing the cost–benefit trade-off. This is a typical situation in large citizen science projects based on human computation, as those hosted at the Zooniverse (zooniverse.org) and the Crowdcrafting (crowdcrafting.org) platforms.

This study analyses *how to automatically improve task replication at execution time by considering participants' credibility scores and the difficulty of tasks*. It focuses on (i) how to automatically measure the difficulty of tasks and the credibility of participants while they execute the tasks in the system, and (ii) how such measures can be used to define, at execution time, the suitable degree of replication for each task. To this end, we go through existing studies on human computation, credibility assessment, and task replication. Based on them, we propose four alternative metrics to measure participants' credibility considering the agreement among themselves. These metrics cover a diversity of participants' features, such as: the amount of generated answers; the amount of agreement with other participants that could be expected to occur through chance alone; and groups of participants that usually generate the most frequent answers. Then, we propose an adaptive task replication algorithm that optimises the degree of replication for each task, taking into account the participants' credibility and the difficulty of the task. The main idea is to stop replication as soon as the algorithm obtains a group of answers that is credible enough. Naturally, there are tasks in which the divergence in the answers is so high that a credible answer is not obtained even by increasing replication. The algorithm is designed to detect these situations, and stop replicating the task when a maximum degree of replication is reached.

Our evaluation study is based on trace-driven simulations [15]. The simulations are guided by data sets collected from two real human computation projects: Sentiment Analysis, and Fact Evaluation. Such data sets comprise hundreds of participants performing thousands of tasks, being valuable sources to analyse the performance of the proposed replication algorithm. In the simulations we evaluate 160 different configurations of the proposed algorithm and also two comparative strategies: (i) an *oracle* that knows whether an answer provided by a human is correct or not, and stops replicating the task when a correct answer is obtained; and (ii) a *majority voting* strategy that collects answers from a fixed number of participants, and identifies as correct the answer provided by the majority of them. We evaluate both the accuracy of answers and the replication reduction reached by these strategies.

The results show that the proposed credibility-based task replication algorithm is effective in achieving replication reduction while meeting other quality of service requirements, such as the required credibility. Some configurations of the algorithm reach higher accuracy than majority voting and achieves a replication reduction comparable with that attained by the oracle. In doing so, it improves the ability of the system to properly use the cognitive power provided by participants, while allows users to address the trade-off between different quality-of-service requirements.

The main contributions of this study are:

- we integrate concepts from four distinct literatures, which are human computation, credibility assessment, inter-rater agreement, and replication of tasks;
- we propose four alternative metrics to automatically measure the credibility of participants while they execute human computation tasks in a system, which are: surface agreement, experienced agreement, weighted agreement, and reputed agreement;
- we propose an adaptive task replication algorithm that optimises the degree of replication of each task according to participants' credibility, task difficulty and quality of service requirements.

These contributions have implications for human computation and related areas that are based on performing tasks with the participation of people, such as the areas of citizen science, crowdsourcing, and social computing. They also have implications for the area of distributed systems. Human computation systems are distributed systems in which computational resources are human beings. As such, some of the concepts employed in traditional distributed systems — i.e. those in which computational resources are machines — to replicate tasks can also be employed to replicate tasks in human computation systems. The study highlights this point, but also puts into perspective new issues in task replication that arise only in human computation systems.

The remainder of this paper is organised as follows. Firstly, we provide background on human computation, credibility assessment, task replication, and also discuss relevant previous work. Next, we present our approach to use agreement-based metrics to assess credibility and replicate tasks in human computation systems. Finally, we evaluate the proposed approach using data from two human computation projects, and then discuss the implications and limitations of the study.

## 2. Background and related work

Now we turn to present the terminology we adopt throughout the paper by briefly reviewing relevant notions of human computation, credibility assessment, and task replication. Thereafter, we discuss the related work.

### 2.1. Background

*Human computation.* Systems based on human computation are distributed systems in which humans participate as computational elements [16,2,3,17]. There are three core entities in this sort of system: requesters, workers, and platforms. *Requesters* are users who act in the system by submitting human computation tasks to be performed. A human computation task (or human intelligence