Future Generation Computer Systems 56 (2016) 526–536

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Workforce-efficient consensus in crowdsourced transcription of biocollections information

Andréa Matsunaga [a,*], Austin Mast [b], José A.B. Fortes [a]

[a] Advanced Computing and Information Systems Laboratory, ECE, University of Florida, 32611, USA
[b] Department of Biological Science, Florida State University, 32306, USA

HIGHLIGHTS

- We describe the challenges faced when trying to reach consensus on data transcribed by different workers.
- We offer consensus algorithms for textual data.
- We implement a consensus-based controller to assign a dynamic number of workers per task and per field of a task.
- We propose the use of clustering to further eliminate redundant work.
- We propose enhancements of future crowdsourcing task assignments in order to minimize the need for complex consensus algorithms.

ARTICLE INFO

Article history:
Received 15 March 2015
Received in revised form
7 July 2015
Accepted 12 July 2015
Available online 20 July 2015

Keywords:
Crowdsourcing
Consensus algorithms
Digitization
Biocollections infrastructure
Biodiversity

ABSTRACT

Crowdsourcing can be a cost-effective method for tackling the problem of digitizing historical biocollections data, and a number of crowdsourcing platforms have been developed to facilitate interaction with the public and to design simple "Human Intelligence Tasks". However, the problem of reaching consensus on the response of the crowd is still challenging for tasks for which a simple majority vote is inadequate. This paper (a) describes the challenges faced when trying to reach consensus on data transcribed by different workers, (b) offers consensus algorithms for textual data, (c) implements a consensus-based controller to assign a dynamic number of workers per task and per field of a task, (d) proposes the use of clustering to further eliminate redundant work and (e) proposes enhancements of future crowdsourcing task assignments in order to minimize the need for complex consensus algorithms. Experiments using the proposed algorithms show multifold increase in the ability to reach consensus when compared to majority voting using exact string matching. In addition, the workforce controller is able to decrease the crowdsourcing cost per task and per task field by 37% and 50%, respectively, when compared to a strategy that uses a fixed number of workers. The accuracy of clustering is also good and it has the potential to increase the quality of tasks that can be clustered.

1. Introduction

Digitization of the scientific information associated with specimens is at full speed in the US biological research collections, as part of the NSF-ADBC (Advancing Digitization of Biodiversity Collections) program [1]. The materials requiring transcription include catalogs, field notes, and especially specimen labels (e.g., labels on specimen sheets or packets, slides, drawers, pins, vials and jars). Fig. 1 shows a sample of the material to be digitized and exposes some of the challenges faced when trying to apply Optical Character Recognition (OCR): handwriting, different types of fonts, faded text, inconsistent positioning of labels, imaging at oblique angles, specimens covering and sometimes hiding text, and the need to interpret or infer information. Crowdsourcing is an appealing solution to lower the cost and accelerate the transcription rate of the large volume of information in these labels which have been amassed over the last two centuries [2]. Several projects [3–7] are proving that volunteers from the general public can perform the transcriptions at speeds that a single project or institution cannot match.

The strategies to set up such a crowdsourcing platform vary, but all of them need to deal with the fact that workers are not always reliable. Given the importance of transcription accuracy for scientific studies, crowdsourced transcriptions require some

* Corresponding author. Tel.: +1 352 392 4964; fax: +1 352 392 5040.
E-mail address: ammatsun@ufl.edu (A. Matsunaga).

http://dx.doi.org/10.1016/j.future.2015.07.004
0167-739X/

Fig. 1. Materials to be digitized: catalogs (center top), and labels next to specimens. Labels appear in jars (top center right), sheets of paper with pressed specimens (top left), vials (bottom left), packets (center left), pinned (center right), hanging from specimens (top right) and in drawers (bottom right).

amount of redundancy. At the same time, there is a trade-off between the level of redundancy (multiple workers transcribing the same material) and the total number of specimen labels transcribed due to limits on available volunteer hours or budgets (when transcribers are compensated). Furthermore, when multiple independent workers transcribe the same material, it is necessary to generate a final consensus amongst all redundant transcriptions. Reaching consensus on a numeric value or on a fixed number of categories is straightforward with majority voting, but textual transcriptions require further effort, especially in the presence of human interpretation of the information contained in specific label fields (leading to non-verbatim transcriptions).

To deal with these challenges in crowdsourcing systems, this paper describes (1) the influence of string comparison algorithms on reaching consensus using transcribed and interpreted data from a popular crowdsourcing platform for digitization of biocollections (NotesFromNature [3]), (2) a strategy to produce a consensus response with excellent accuracy, (3) a controller that minimizes the number of workers required for a particular task and for each field of a task, (4) use of clustering to further minimize the workforce requirement and improve the quality of certain tasks, and (5) insights to improve the design of crowdsourcing tasks.

This paper is organized as follows: Section 2 presents related work, Section 3 details our proposed consensus algorithms and worker controller, Section 4 shows our experimental results, Section 5 presents improvements for future work, and Section 6 concludes this work.

2. Related work

Several projects and software tools have started to address the particular scenario of biological collection transcription. Some take advantage of OCR to generate as much transcription as possible, and Natural Language Processing (NLP) to parse information from OCR into appropriate fields. Noteworthy efforts are shown in Table 1. The workflow strategies of these projects vary greatly: NotesFromNature [3] requires a fixed number of transcriptions to consider a specimen digitization task complete; Atlas of Living Australia's DigiVol (ALA DigiVol) [4] and Symbiota [5] require a single transcription; ScioTR [6] produces alternatives from OCR and NLP, from which the user selects the best option; and Transcribe Bentham [7,8] makes use of multiple collaborative volunteers. The need for consensus is either avoided (where there are single or collaborative transcribers), or is turned into another crowdsourcing task, or is simply absent with multiple answers returned to the task requester. General purpose crowdsourcing platforms, such as Amazon Mechanical Turk [9] and PyBossa [10], have enabled a number of studies where the tasks required only a simple majority voting while surveying users [11], categorizing content [12,13], assessing translation [14,15], among numerous other micro-tasks. Enki [16] is an analysis tool for generating simple statistics (e.g., counts, unique values, and averages) on crowdsourcing tasks produced through PyBossa, and it does not include consensus algorithms for textual data. Thus, our contributions complement all these crowdsourcing platforms by increasing the confidence in accuracy with use of multiple workers and offering consensus on multiple answers.

Data quality on crowdsourced tasks has been widely studied by analyzing how human factors influence responses [12], offering methods to separate error from bias [13], comparing responses from the crowd with those provided by experts [14,15,17], detecting and filtering noisy workers based on their z-score [18] or a support-vector-machine algorithm [19], analyzing the use of a small amount of expert data to supervise the crowd responses [20], and proofing the response of workers with another crowdsourcing task [21]. In this work, we also assess the quality of the consensus output by comparing with transcriptions performed by an expert.

Minimizing the amount of crowdsourcing work, and consequently its cost, has mainly been tackled by the addition of a phase to validate the responses from previous volunteers by experts [4,5,7] or other workers [19,21]. However, this puts pressure on experts and/or more weight on the quality of workers validating the data. In this work, we propose to control the number of workers on a per-task basis, based on the consensus reached by responses from independent workers. The main idea behind the controller is that the number of transcriptions needed to reach consensus will depend on the difficulty and amount of interpretation required by the task and its sub-parts, while providing a threshold level of accuracy. TurKit [22] is a general purpose toolkit for interactive programming against Amazon Mechanical Turk platform, and it can be