Quality Control for Crowdsourced Enumeration Tasks

Shunsuke Kajimura
The University of Tokyo

Yukino Baba
National Institute of Informatics

Hiroshi Kajino
The University of Tokyo

Hisashi Kashima
Kyoto University

Abstract

Quality control is one of the central issues in crowdsourcing research. In this paper, we consider a quality control problem of crowdsourced enumeration tasks that request workers to enumerate possible answers as many as possible. Since workers neither necessarily provide correct answers nor provide exactly the same answers even if the answers indicate the same idea, we propose a two-stage quality control method consisting of the answer clustering stage and the reliability estimation stage.

1 Introduction

In this paper, we focus on enumeration tasks that request workers to enumerate possible answers as many as possible. An example of enumeration tasks is finding the names of islands in the Great Barrer Reaf. There are two types of uncertainties in the enumeration tasks; workers neither necessarily provide answers correctly satisfying the requirement of the given task nor provide exactly the same answers for a specific object due to the variations of texts and numerical values. Therefore, quality control plays a more significant role in crowdsourced enumeration. Trushkowsky et al. (2013) considered the estimation problem of the number of items of a enumeration task; however, to the best of our knowledge, there has not been proposed any quality control method applicable to enumeration tasks.

We aim to resolve the uncertainties in crowdsourced enumeration tasks by selecting a subset of the answers each of which indicates a different object and correctly satisfies the requirement of the given task. We introduce a two-stage quality control method for enumeration tasks consisting of an *answer clustering* stage followed by a *reliability estimation stage*.

2 Problem Setting

Let us assume that there are W workers, let $\mathcal{T}_i$ denote a set of answers given by the i-th worker, and let N denote the number of answers provided by all the workers. Note that we usually expect workers not to provide multiple answers indicating the same object by explicit or implicit instructions. We denote the set of all answers provided by workers by $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2 \cup \cdots \cup \mathcal{T}_W$. Further, we assume that we have the non-negative dissimilarity d_{uv} between every pair of answers u and v in $\mathcal{T}$.

Given the answer set, $\{\mathcal{T}_i\}_{i=1,\dots,W}$, and $D = \{d_{uv}\}_{u,v=1,\dots,N}$, our goal is to obtain a subset of given answers, $\mathcal{P} \subseteq \mathcal{T}$, so that each answer in the answer subset correctly satisfies the requirement of the given task and indicates a different object.

3 Two-Stage Quality Control Method for Enumeration Tasks

3.1 Answer Clustering

In the first stage, all of the answers are clustered so that the answers in each cluster indicate the same object. Each cluster is represented by a single answer, which we call a *representative answer*. This answer clustering stage aims to resolve the uncertainty of noisy answers indicating the same object. We resort to apply an exemplar clustering method. In addition, our instructions that workers should not answer multiple answers indicating the same object naturally lead us to add a *cannot-link* constraint to an exemplar clustering (Elhamifar, Sapiro, and Vidal 2012). We formalize a constrained exemplar clustering by incorporating cannot-link constraints to an exemplar clustering method which is formalized as a convex optimization problem by Elhamifar, Sapiro, and Vidal (2012).

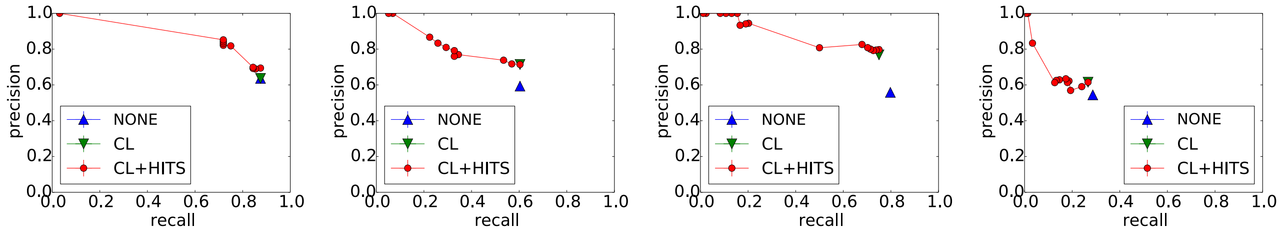
Let $z_{uv} \in [0, 1]$ denote the probability that the answer u is represented by the answer v, and $Z = \{z_{uv}\}_{u,v=1,\dots,N}$ denote the representative matrix. Given the dissimilarity matrix D, the exemplar clustering problem is defined as

$$\min_{z_{uv}} \quad \sum_{v=1}^{N}\sum_{u=1}^{N} d_{uv} z_{uv} + \lambda \sum_{u=1}^{N} ||\boldsymbol{z}_{u,:}||_q \tag{1}$$

$$\text{s.t.} \quad \sum_{u=1}^{N} z_{uv} = 1,\ \forall v, \quad z_{uv} \geq 0,\ \forall u, v, \tag{2}$$

$$\max_{u} \sum_{v \in T_i} z_{u,v} = \left\| \sum_{v \in T_i} \boldsymbol{z}_{:,v} \right\|_{\infty} \leq 1,\ \forall i, \tag{3}$$

where $\boldsymbol{z}_{u,:}$ denotes the u-th row of Z, $z_{:,v}$ denotes the v-th column of Z, $\lambda > 0$ is a regularization parameter which



(a) Task 1 (b) Task 2 (c) Task 3 (d) Task 4

Figure 1: Comparison of the precision-recall curves. In most cases, our reliability estimation stage improved the precision of the answers.

controls the number of clusters, and $q \in \{2, \infty\}$. Elhamifar, Sapiro, and Vidal (2012) gave the formulation of exemplar clustering (1) and (2), and we append cannot-link constraints (3).

3.2 Reliability Estimation

The second stage estimates the reliabilities of representative answers. Based on them, representative answers of low reliability can be removed.

We define a representative answer as answer u such that $\|z_{u,:}\| \neq 0$, which represents the cluster. Let $\mathcal{U} = \{u \in \mathcal{T} \mid \|\boldsymbol{z}_{u,:}\| \neq 0\}$ denote a set of representative answers. To select correct answers, we next estimate the reliabilities of the representative answers. We assume that an answer is reliable if multiple reliable workers provide it, and a worker is most likely to be reliable if the person provides many reliable answers. This notion is similar to the one employed in the HITS algorithm (Kleinberg 1998).We apply the HITS to the answer reliability estimation with analogies between authorities and answers, and between hubs and workers.

4 Experiments

We applied our two-stage quality control method to several POI collection tasks that we posted to Lancers, a crowdsourcing service provided in Japan. A POI collection task asks workers to enumerate the longitude and latitude of points that satisfy a given requirement of the task. We prepared four POI collection tasks that aim to collect locations of telephone booths around Shimbashi station (Task 1), locations of noodle restaurants around Takamatsu station (Task 2), locations of mail boxes around Ueno station (Task 3), and locations of public toilets around Shinjuku station (Task 4). The numbers of answers, workers, and answers in ground truth of Task 1, 2, 3, and 4 are (133,4,96), (63,7,58), (122,14,84), and (82, 11, 150), respectively.

We used precision and recall as our evaluation metrics. Let v denote an answer represented as a vector, $(\mathrm{longitude}, \mathrm{latitude})$. We considered that each obtained the answer $v \in \mathcal{P}$ is correct if the distance between the answer v and the nearest ground truth is smaller than the threshold value $d = 0.0003$.

We compared the quality of the answers produced by our method (called "CL+HITS") with the quality of ones without our quality control method (called "NONE"), and of ones that without the reliability estimation stage (called "CL") to verify the effectiveness of each stage of our two-stage method.The parameters in the answer clustering stage were set as $q = \infty$ for solving the problem by a linear programming and $\lambda = 0.1$.

We show comparisons of precision-recall curves between the methods are shown in Figure1. In all the tasks except Task 4, our two-stage method improved precision as recall becomes lower. This means our two-stage method allows us to control a balance between precision and recall. The reliability estimation stage successfully filtered out the spam workers to obtain preferable answers in the Task 1 and 3. However, the workers showed high accuracies in the Task 2, and consequently, the reliability estimation stage did not make any change in the results. However, our two-stage method did not improve precision as recall becomes lower in Task 4. This result was caused by presence of "streaker", a worker who gives a much larger number of answers (Trushkowsky et al. 2013), and by the fact that the accuracy of the answers provided by the streaker was inferior in Task 4. Since a part of the answers given by the worker was overlapped with the ones provided by the other trustworthy workers, our reliability estimation stage assigned a high reliability to the worker. Therefore, the answers of the worker were considered as reliable. Such an undesirable situation is difficult to resolve because there is no information to judge whether an answer which is not overlapped with others is correct or not.

References

Elhamifar, E.; Sapiro, G.; and Vidal, R. 2012. Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. In *Advances in NIPS*.

Kleinberg, J. 1998. Authoritative sources in a hyperlinked environment. In *Proceedings of SODA*.

Trushkowsky, B.; Kraska, T.; Franklin, M. J.; and Sarkar, P. 2013. Crowdsourced enumeration queries. In *Proceedings of ICDE*.