Consensus-based Techniques for Range-task Resolution in Crowdsourcing Systems

Lorenzo Genta
Dipartimento di Informatica
Università degli Studi di Milano
Via Comelico 39
20135 - Milano, Italy
genta@di.unimi.it

Alfio Ferrara
Dipartimento di Informatica
Università degli Studi di Milano
Via Comelico 39
20135 - Milano, Italy
ferrara@di.unimi.it

Stefano Montanelli
Dipartimento di Informatica
Università degli Studi di Milano
Via Comelico 39
20135 - Milano, Italy
montanelli@di.unimi.it

ABSTRACT

In crowdsourcing, a range task is a type of creation task where only free answers belonging to the numeric domain are accepted/possible. In this paper, we present the *median-on-agreement* (ma) techniques based on statistical and consensus-based mechanisms for determining the result of range tasks. The ma techniques are characterized by i) the distinction between the group of workers that agree (i.e., workers in the consensus) on the task result from the group that disagree, and ii) the calculation of the final task answer through a median-based mechanism where only answers of workers in the consensus are considered.

Keywords

crowdsourcing, consensus evaluation, range task management

1. INTRODUCTION

In the recent years, crowdsourcing systems have gained growing popularity as powerful solutions for addressing the execution of complex, time-consuming activities where the contribution of human workers can be decisive and the use of automatic procedures is not completely effective, such as for example collaborative filtering and web-resource tagging. Usually, in this kind of systems, crowd workers are involved in *decision* tasks where they are called to select the most appropriate answer among a set of predefined alternatives (e.g., [9]). In a conventional scenario, multiple workers participate to the execution of a task, thus multiple answers are collected and the final result is derived by assessing the level of agreement between the different answers and by deciding if a consensus has been reached [1, 3]. The use of crowdsourcing systems is now being proposed also for the resolution of the so-called *creation* tasks, in which the task answer can be any kind of worker-generated content like for example a free text answer as well as a drawing or another visual/multimedia artifact. This task type enables the worker to express her/his creativity, thus enabling crowdsourcing to become a mechanism for collaborative knowledge creation. However, in creation tasks, the problem of choosing the final task result among all the available worker answers is even more challenging than for decision tasks, especially when the task question is intrinsically subjective and a factual answer is not possible nor appropriate (e.g., a labeling task in which the worker is called to provide a featuring keyword for a group of web images).

In this paper, we focus on range tasks, namely a type of creation task where only free answers belonging to the numeric domain are accepted/possible [1]. We propose the *median-on-agreement* (ma) techniques based on statistical and consensus-based mechanisms. In particular, the ma techniques are conceived to address range task resolution when multiple crowd workers are involved in the execution of each task. Each worker autonomously and independently executes a task, thus a number of different answers is produced. Based on these answers, the ma techniques allow i) to distinguish the group of workers that agree (i.e., workers in the consensus) on the task result from the group that disagree, and ii) to calculate the final task answer through a median-based mechanism where only answers of workers in the consensus are considered. The application of the ma techniques to the Argo crowdsourcing system is presented as well as experimental results against the main state-of-the-art approaches for range task resolution.

The paper is organized as follows. In Section 2, we illustrate motivations and related work. The ma techniques are presented in Section 3. In Section 4, the application of ma to Argo is discussed. In Section 5, experimental results on a real crowdsourcing case-study are presented. Concluding remarks are provided in Section 6.

2. MOTIVATING SCENARIO

Consider the scenario described in [6] where the use of a crowdsourcing approach is proposed for estimating the amount of calories in a meal. In [6], a task is characterized by a picture of a dish and a worker receiving a task to execute is asked to insert a numeric value corresponding to her/his calorie estimation based on the given picture.

This is an example of a range task, in that a worker receiving a task to execute can only provide a free numeric answer, namely integer or decimal value, based on her/his personal point-of-view, knowledge, perception, and expertise. This means that no predefined options/suggestions are available and workers are called to independently and au-

tonomously provide her/his own task answer. Moreover, the real amount of calories in a dish (i.e., in a task) is not available/known and only a *collective answer* is possible [3]. This means that crowdsourcing has the goal to provide a result that represents the so-called "wisdom of the crowd", in which the reliability of a task result is determined by its credibility: the more the consensus among workers on an answer is high, the more the answer reliability is high.

An intuitive and popular solution for range task resolution is to employ a mean-based approach in which multiple workers are involved in the execution of each task and the arithmetic mean of the whole set of worker answers is provided as final result [5]. The main drawbacks of a mean-based approach are illustrated by Francis Galton in [4] where the use of arithmetic mean for computing the result of a range task is deprecated, since it

> *would give a voting power to "cranks" in proportion to their crankiness. One absurdly large or small estimate would leave a greater impress on the result than one of reasonable amount, and the more an estimate diverges from the bulk of the rest, the more influence would it exert.*

In other words, the numeric answer of a single worker that diverges (i.e., it is very different) from the other more-or-less equivalent worker answers has a strong influence on the final task result. This means that a single worker can autodetermine her/his impact on the task result independently from her/his trustworthiness. This is especially true when the group of workers involved in a task execution is small (i.e., 5-10 workers per group) and malicious or inaccurate workers can be involved as usually occurs in real systems.

Further work on resolution of range tasks are presented in [7]. This contribution is in the field of QoE (Quality of Experience) where workers are asked to provide an evaluation of their experience with a service (e.g., web browsing, phone call, TV broadcast). The authors propose a technique called CrowdMOS (i.e., *Crowd*sourcing *M*ean *O*pinion *S*core) based on the analysis of the answer distribution provided by workers. The high subjectivity/uncertainty of considered tasks motivates the use of a random-effects model for determining the task result. However, only random variables based on a normal distribution (i.e., a symmetric distribution) can be used for representing errors, thus other statistical distributions are not supported.

In the following, we propose consensus-based techniques for managing range task resolution based on two main contributions. First, use of the median value (instead of the arithmetic mean) to determine the task result which is representative of the multiple answers collected from the involved workers. Second, use of consensus as a mechanism for distinguishing workers that agree on the task result from workers that disagree and represent a sort of outlier position.

3. THE MEDIAN-ON-AGREEMENT TECHNIQUES

Consider a range task T assigned to a group of workers $G = \{w_1, \ldots, w_n\}$ providing a set of answers $A = \{a_1, \ldots, a_n\}$ where $a_k \in A$ is the numeric answer provided by the worker $w_k \in G$. Range task resolution according to the ma techniques is articulated in two main steps: *identification of the support group* and *definition of the final task result* described in the following.

Identification of the support group. We call $G_{CA^1} \subseteq G$ the *support group* of G, namely the group of workers that agree on the task result. Two workers agree on the task result when they provide a similar numeric answer, meaning that the values provided in the task answer are near in comparison with the overall range of answers A provided by all the workers in G. We call $A_{CA^1} \subseteq A$ the set of task answers provided by the workers in G_{CA^1}. Consider the median value m_A of all the provided worker answers A. The group G_{CA^1} is progressively built by including workers that provided an answer close to m_A, namely:

1. Compute the median m_A over the whole set of worker answers A and define $G_{CA^1} = \emptyset$, $A_{CA^1} = \emptyset$.

2. Select the worker answer $a_k \in A$ which is nearest to m_A. Insert a_k in A_{CA^1} and insert the worker w_k in the support group G_{CA^1}.

3. The *coefficient of variation cv* is exploited to decide whether an answer $a_k \in A$ is near enough to m_A for being included in G_{CA^1}. To this end, cv is calculated over the set of answers in A_{CA^1}:

$$cv(A_{CA^1}) = \frac{\sqrt{\frac{1}{|A_{CA^1}|}\sum_{i=1}^{|A_{CA^1}|}\left(a_i - \mu_{A_{CA^1}}\right)^2}}{\mu_{A_{CA^1}}}$$

   where $|A_{CA^1}|$ is the number of answers in A_{CA^1}, a_i represents the i^{th} worker answer in A_{CA^1}, and $\mu_{A_{CA^1}}$ represents the arithmetic mean of the answers in A_{CA^1}.

4. The insertion of workers in G_{CA^1} is repeated until the coefficient of variation over the answers A_{CA^1} is lower than a threshold th_{cv} (i.e., go back to step 2 if $cv(A_{CA^1}) < th_{cv}$). Otherwise, remove the last-inserted item from G_{CA^1} and A_{CA^1} and continue with the next step.

5. Create the set $G_{CA^2} = G \setminus G_{CA^1}$ containing the workers that are not in the support group. Analogously, the set $A_{CA^2} = A \setminus A_{CA^1}$ is created as well.

Definition of the final task result. The final task result $\bar{A}$ is defined as the median value calculated over the set of worker answers A_{CA^1}, namely $\bar{A} = m_{A_{CA^1}}$.

Example. Consider a task T_1 where workers are asked to guess the distance between the two Italian cities Caserta and Siena in kilometers (the real distance is 352 Km). Consider the following set of worker answers: $A = \{300, 300, 301, 301, 350, 351, 351, 351, 351, 400, 408, 408, 450, 500, 600, 600, 600, 650, 700, 1500\}$. The median value over the whole set of worker answers $m_A = 404$. According to ma, we consider a threshold for the coefficient of variation $th_{cv} = 0.15$ and we identify the support group G_{CA^1} shown in Figure 1. With this support group, the median value of the answers provided by workers in the support group is returned as final task result: $\bar{A} = m_{A_{CA^1}} = 351$.

4. APPLICATION TO THE ARGO SYSTEM

The ma techniques have been implemented in the Argo crowdsourcing platform (http://island.ricerca.di.unimi.it/projects/

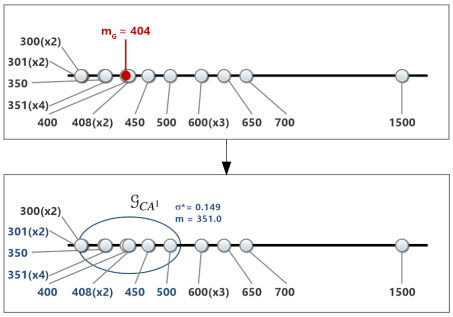

Figure 1: Identification of support group in ma

argo/ (Italian language)). In Argo, range task resolution is enforced through *consensus-based evaluation techniques* and *trustworthiness-based worker management* by relying on our experience and research results in this field [3].

Consensus-based evaluation of range tasks. For consensus evaluation, Argo employs a *weighted-voting mechanism* called **supermajority** where the answer of a worker w_k has a weight corresponding to her/his trustworthiness. Supermajority is based on the verification of two different constraints called *quorum-constraint* (q) and *balance-of-power constraint* (bop). The q-constraint verifies that the task result $\bar{A}$ is supported by a group of workers G_{CA^1} with enough weight (i.e., trustworthiness) for satisfying a given quorum $q \in [0.51, 1]$. The bop-constraint verifies that a single worker cannot shift the majority from one answer to another one just by changing her own task answer [3]. This means that the support group G_{CA^1} still satisfies the q-constraint even if a worker is shifted from G_{CA^1} to G_{CA^2}. A task is committed on the task result $\bar{A}$ when the supermajority constraints are satisfied (i.e., consensus is verified). On the opposite, when supermajority constraints are not satisfied, the task remains uncommitted. In this case, the task should be re-executed or considered as failed.

Trustworthiness-based worker management. The Argo system aims at keeping into account not only the mere effort workers spent in executing tasks, but also the quality of the effort provided. A worker W is characterized by a *worker score* σ_W, and a *worker trustworthiness* τ_W.

The worker score σ_W represents the worker revenue composed by i) a salary, the payment the worker receives each time she/he executes a task, regardless of the consensus verification, and ii) an award, a bonus the worker receives each time she/he contributes to commit a task.

The worker trustworthiness $\tau_W \in [0, 1]$ is defined to capture the worker ability to foster the task commitment and it is based on the worker history in executing tasks. At the beginning of the crowdsourcing activities (time $t = 0$), the worker trustworthiness τ_W is set to an initial value $\tau_W^0 = \tau_0$. Each time a task T is committed (time $t+1$), the trustworthiness of a worker $W \in G$ is updated. In particular, the worker trustworthiness increases (i.e., $\tau_W^{t+1} > \tau_W^t$) when the worker belongs to the support group (i.e., $W \in G_{CA^1}$), thus confirming her/his ability to foster task commitment in the last-executed task T. On the opposite, the worker trustworthiness decreases when the worker is not in the support group (i.e., $W \notin G_{CA^1}$).

5. EXPERIMENTAL RESULTS

For evaluation of the proposed ma techniques, we consider the geo-dis case-study for crowdsourcing the geographic distance between pairs of Italian cities.

The experiment has been executed by relying on the Argo prototype. We collected a dataset of 120 Italian cities with their geographic coordinates extracted from the FreeBase (http://www.freebase.com) open repository. We built a set of 634 tasks each one asking for the distance between a pair of different cities. The experimentation on geo-dis was conducted with a crowd of 585 workers selected in a class of master-degree students (average worker age is 21 years old). For task resolution, we asked the workers to rely on their personal knowledge and we set the allowed time to perform a task to a maximum of 15 minutes. In the experimentation, the Argo prototype has been configured as follows: i) initial worker trustworthiness $\tau_0 = 0.5$; ii) group size s_G=20; iii) quorum value $q = 0.51$; iv) the worker salary is $s = 0.1$ and the worker award is $a = 1$.

Evaluation is based on two different experiments over the geo-dis case-study. The former experiment presents a comparison of the ma techniques implemented in the Argo system (ma_{Argo}) against other state-of-the-art techniques for range task resolution. The latter experiment is performed to evaluate the crowdsourcing cost of the ma techniques by measuring the number of committed/uncommitted tasks.

Comparison against state-of-the-art techniques. We compare ma_{Argo} against the following competitor techniques:

Overall arithmetic mean μ_O. This method refers to the classical approach proposed in [10] where the result of a task T is given by computing the arithmetic mean over all the obtained answers.

Outlier-cleaned arithmetic mean - Standard Deviation μ_{2SD}. This method consists in applying a classical outlier removal technique based on the standard deviation (2SD) [8] to the set of answers of a task T. After removal of the outliers, the arithmetic mean is finally computed over the remaining answers.

Outlier-cleaned arithmetic mean - Median Rule μ_{MR}. This method consists in applying a more recent outlier removal technique based on the median rule [2] to the set of answers of a task T. After removal of the outliers, the arithmetic mean is finally computed over the remaining answers.

Overall median m_O. This method consists in computing the result of a task T as the median value of all the provided answers. As far as we know, state-of-the-art techniques based on the median value are not provided. However, we compare ma_{Argo} against m_O since this is the natural baseline for our ma techniques.

In the evaluation, we consider ma_{Argo} under three configurations characterized by different thresholds for the coefficient of variation th_{cv}. Results are evaluated through *average-error* and *average-error with outlier-removal* mechanisms. In the average error mechanism, for each task T, the evaluation considers the error between the distance estimation in the crowdsourcing result $\bar{A}$ and the real distance between the two cities contained in T. The average error $\bar{\epsilon}_A$ is calculated as:

$$\bar{\epsilon}_A = \frac{\sum_{i=1}^{i=n} |A_i - R_i|}{|T|}$$

where $n = |T|$ is the overall number of tasks, A_i is the crowd-

sourcing result of the task T_i, R_i is the real distance between the pair of cities in the task T_i calculated through the geodesic distance. In the average-error with outlier-removal, the error evaluation follows the same approach of $\bar{\epsilon}_A$ calculation, but outliers are removed according to the conventional criterion based on standard-deviation (2SD) [8].

The results of this experiment are presented in Table 1. The first consideration that has to be done is about the result of the μ_O technique. The fact that the obtained average error $\bar{\epsilon}$ is so high is mainly due to the presence of malicious workers in a very high number of groups. These malicious workers gave completely wrong answers (e.g., 10 millions kilometers as distance between Rome and Milan) that have a very serious impact on the task result when the arithmetic mean is considered and outlier removal is not performed. We note that the median-based techniques (i.e., m_O and ma_{Argo}) provide better results than the techniques based on the standard deviation. We argue that this is due to the assumption of symmetric distribution used in μ_O, μ_{2SD}, and μ_{MR}, which is is usually false (e.g., see for example the task presented in Figure 1). As a general remark, we observe that the median-based solutions provide better results than mean-based techniques even without the outlier-removal phase. By considering the ma_{Argo} results with the different thresholds on the coefficient of variation, we note that the lower is the threshold th_{cv}, the lower is the average error $\bar{\epsilon}$. This means that a more restrictive mechanism for determining the support group G_{CA^1} increases the accuracy of obtained results.

Table 1: Comparison of results

	$\bar{\epsilon}$ (Km)	$\bar{\epsilon}^c$ (Km)
μ_O	666206.10	9558.12
μ_{2SD}	48.44	40.98
μ_{MR}	19.71	14.00
m_O	18.10	11.21
ma_{Argo} $(th_{cv} = 0.25)$	12.89	6.71
ma_{Argo} $(th_{cv} = 0.15)$	9.15	5.18
ma_{Argo} $(th_{cv} = 0.05)$	2.69	1.35

Analysis on the task commitment We observed that a low value of th_{cv} produces a low average error $\bar{\epsilon}$. However, on the opposite, a low value of th_{cv} also produces a high number of uncommitted tasks, and thus high expenses for crowdsourcing execution. For this reason, in this experiment, we analyze the number of committed tasks when different thresholds on the coefficient of variation are considered. To this end, we define the *commitment ratio* as follows:

$$c = \frac{N_c}{N_c + N_u}$$

where N_c is the number of committed tasks and N_u is the number of uncommitted tasks.

The commitment ratio for different coefficient of variation thresholds th_{cv} are presented in Table 2. We note that the lower is the coefficient of variation threshold, the lower is the value of commitment. This behavior is motivated by the fact that the lower is the coefficient of variation, the more restrictive is the mechanism for determining the support group G_{CA^1}. As a result, it is important to configure the crowdsourcing execution by tuning the threshold th_{cv} with the goal to set the desired tradeoff between accuracy of results and commitment ratio. In the geo-dis case study, the threshold value $th_{cv} = 0.15$ provides the best tradeoff between accuracy (i.e., almost twice value on accuracy with respect to the other threshold values) and commitment ratio (i.e., $c \approx 90$ %).

Table 2: Commitment evaluation

	#Committed	c
$th_{cv} = 0.25$	624	**98.4%**
$th_{cv} = 0.20$	609	**96.1%**
$th_{cv} = 0.15$	565	**89.1%**
$th_{cv} = 0.10$	507	**80.0%**
$th_{cv} = 0.05$	422	**66.6%**

6. CONCLUDING REMARKS

In this paper, we presented the ma techniques for range task resolution in crowdsourcing systems. Application to the Argo system as well as experimental results on a real case-study are provided to show the contribution of the proposed solution with respect to the state-of-the-art. Ongoing work are focused on the so-called *task routing problem* with the goal to specify a family of configuration patterns for dynamically choosing the most appropriate group of workers that can be selected for assignment of a given task to be executed based on worker expertise and knowledge.

7. REFERENCES

[1] A. Bozzon, M. Brambilla, S. Ceri, and A. Mauri. Reactive Crowdsourcing. In *Proc. of the 22nd Int. World Wide Web Conference (WWW 2013)*, pages 153–164, Rio de Janeiro, Brazil, 2013.

[2] K. Carling. Resistant Outlier Rules and the Non-Gaussian Case. *Computational Statistics & Data Analysis*, 33(3):249–258, 2000.

[3] S. Castano, A. Ferrara, L. Genta, and S. Montanelli. Combining Crowd Consensus and User Trustworthiness for Managing Collective Tasks. *Future Generation Computer Systems*, 54, 2016.

[4] F. Galton. One Vote, One Value. *Nature*, 75:414, 1907.

[5] T. W. Malone, R. Laubacher, and C. Dellarocas. The Collective Intelligence Genome. *IEEE Engineering Management Review*, 38(3), 2010.

[6] J. Noronha, E. Hysen, H. Zhang, and K. Z. Gajos. Platemate: Crowdsourcing Nutritional Analysis from Food Photographs. In *Proc. of the 24th symposium on User Interface Software and Technology*, pages 1–12, Santa Barbara, CA, USA, 2011.

[7] F. P. Ribeiro, D. A. F. Florêncio, C. Zhang, and M. L. Seltzer. CROWDMOS: An Approach for Crowdsourcing Mean Opinion Score Studies. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2416–2419, Prague, Czech Republic, 2011.

[8] S. Seo. *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*. PhD thesis, University of Pittsburgh, Pennsylvania, USA, 2006.

[9] C. Sun, N. Rampalli, F. Yang, and A. Doan. Chimera: Large-scale Classification Using Machine Learning, Rules, and Crowdsourcing. *Proceedings of the VLDB Endowment*, 7(13), 2014.

[10] J. Surowiecki. *The Wisdom of Crowds*. Random House LLC, 2005.