Worker Ranking Determination in Crowdsourcing Platforms using Aggregation Functions

David Sánchez-Charles, Jordi Nin, Marc Solé and Victor Muntés-Mulero

Abstract— **The increasing adoption of crowdsourcing for commercial and industrial purposes rises the need for creating sophisticated mechanisms in crowd-based digital platforms for efficient worker management. One of the main challenges in this area is worker motivation and skill set control and its impact on the output quality. The quality delivered by the workers in the crowd depends on different aspects such as their skills, experience, commitment, etc. The lack of generic and detailed proposals to incentive workers and the need for creating ad-hoc solutions depending on the domain make it difficult to evaluate the best rewarding functions in each scenario. In this paper, we make a step further in this direction and propose the use of aggregation functions to evaluate the professional skills of crowd-workers based on the quality of their past tasks. Additionally, we present a real industrial crowdsourcing solution for software localisation in which the proposed solutions are put into practice with real text translations quality measures.**

I. Introduction

THE generalisation of on-line social networks, the increase in the unemployment rates in many countries because of the economic recession and the increasing need of industry to flexibly involve human beings in big data analytics and processing are just three important motivations for the growth of the number of platforms and solutions using crowdsourcing strategies. By using these strategies, complex problems can be solved through the use of unparalleled mechanisms to allow for the collaboration of thousands of remote Internet users to solve a specific task, leveraging the potential of emerging intelligence.

Trends seem to be pointing to this model as gaining the position to complement cloud computing: connecting people and machines in a single network. Nowadays, millions of people are asynchronously analysing, synthesising, providing opinion and labelling and transcribing data that can be automatically mined, indexed and even learned. Human brain-guided computation is able to perform tasks that computers can hardly do, at overwhelming speeds. Tagging a picture or a video based on their content or answering questions in natural language, are just a couple of examples.

Besides, the current economic recession affects millions of families worldwide. Just as an example, the unemployment rate in Spain and Portugal was 26.7 and 15.5, respectively (November 2013)[1]. With this situation, unemployment is not only restricted to workers with low levels of education, but it also affects highly qualified professionals. Apparently, this situation is just marking the beginning of a long depression period and it may be a catalyst for crowdsourcing to consolidate as a common new mechanism for outsourcing.

Different authors try to tackle one of the main challenges in crowdsourcing: output quality [1], [6], [9], [11], [22], [28]. Quality is in general linked to the incentive mechanisms used in a specific platform [8]. A clear motivation may potentially lead to higher commitment and better quality. There may exist many different types of incentives including financial, social and moral incentives. Incentives may also be extrinsic, such as money or social approval, or intrinsic, such as fun, knowledge or moral satisfaction. Different mechanisms may also respond to different purposes such as growing the community, increasing the speed or quality in task resolution or retaining workers. In general, industrial applications pursue lucrative objectives. Because of this, industrial applications based on the crowd are quite more constrained in terms of motivating the crowd and tend to reward workers economically. The work presented in [19], for instance, confirms the importance of money compared to other motivations in certain cases. From the industrial perspective, first steps have been done to establish the basis for crowd coordination and create rewarding mechanisms that are based on involving human beings in the evaluation of the quality of other workers through the so-called AV-Units [13]. To the best of our knowledge, there are not generic mechanisms in order to evaluate quality and reward workers in the crowd, that can be adapted to different context and requirements. Besides, complex tasks may require to classify workers in different categories and reward them proportionally to the value they provide to the whole resolution process.

In this paper, following the proposal in [13], we assume that human beings are required to evaluate the quality of the output of a task, as an indirect mechanism to evaluate the skills of other workers in the crowd [5], [10]. We propose to use the already calculated quality evaluation of past tasks as the input of an aggregation function. The aggregation function fuses them in a single datum able to summarise, rank and determine the profile and skills of individuals. Taking into account such profiles, we can better determine who are the most suitable individuals for a given task and what is the

David Sánchez-Charles and Victor Muntés-Mulero are researchers at CA Labs, CA Technologies, in Barcelona (Spain). {David.Sanchez, Victor.Muntes}@ca.com

Jordi Nin is with the Barcelona Supercomputing Center (BSC), Universitat Politécnica de Catalunya - BarcelonaTech, in Barcelona (Spain). nin@ac.upc.edu

Marc Solé is with the Universitat Politécnica de Catalunya - BarcelonaTech, in Barcelona (Spain). msole@ac.upc.edu

This work is partially supported by the Ministry of Science and Technology of Spain under contract TIN2012-34557 and by the BSC-CNS Severo Ochoa program (SEV-2011-00067)

[1] ”Seasonally adjusted unemployment”. Eurostat. http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/3-08012014-BP/EN/3-08012014-BP-EN.PDF

fairest economical reward. Note that, aggregation functions provide us a huge flexibility in the summarisation process, allowing us to define rewards and punishments within their weighting vector. Additionally to that, with our system it is also possible to establish an automatic promotion system for crowd workers based on their past tasks evaluations. This fact may also increase the workers' commitment for delivering high quality tasks because their quality will determine their future rewards.

This paper is organised as follows. Section II introduces previous work including some basic properties of aggregation functions. Then, Section III describes our approach for worker Ranking automatic evaluation. Section IV presents an example of a crowdsourcing platform developed at CA Technologies[2] for crowd-based text translation. Finally, Section V concludes and draws some future research lines.

II. Previous Work

Crowdsourcing [7], [3] is the practice of obtaining services, ideas, or content by requesting contributions from a large online community of people, rather than from traditional employees or providers. For instance, Amazon Mechanical Turk (mturk.com) is a crowdsourcing marketplace that enables companies or individuals to utilise human intelligence to perform tasks that are difficult for computers. Other examples like CrowdFlower (crowdflower.com) or ClickWorker (clickworker.com), extend Mechanical Turk capabilities offering a variety of crowdsourcing services. They improve quality by using gold standard units, redundant reviews of each data unit, etc. Their workflow management system divides complex tasks into smaller units and distributes them among the crowd based on the profile of individuals.

A. Quality Assurance in Crowdsourcing platforms

Venetis and Garcia-Molina propose "Gold Standard Performance" to detect workers' performance before the crowdsourcing task starts [22]. Workers' characteristics such as demographics or personality traits are related to the quality of their work under specific task conditions [9]. The worker perception of five quality assurance mechanisms is also studied in [18]. In general, it is considered that inaccurate acceptance or rejection may not only affect a specific task in a platform, but may also encourage other fraudsters to misbehave in the platform. For example, Hirth et al. raise "Majority Decision Approach" [6] to judge whether worker's submission is correct in simple tasks, and using "Control Group Approach" method in complicated cases. Crowdsource the quality evaluation of the jobs performed by the crowd to avoid the use of such gold standard units, has been proposed as an alternative. The main idea behind this proposal is that human beings are the best quality evaluation method [5], [10].

There are also many other discussion on quality control for crowdsourcing in certain fields such as real-time applications [11]. The debate on the relationship between task quality and rewards has been analysed in [8]. For some scenarios, this relationship has been studied not to be relevant. For instance, in 2009, Yahoo's research institute made a quantitative analysis on the relationship between "Financial Incentives" and "Performance of Crowds" [12], and found that higher rewards can accelerate the accomplishment of the task, but cannot improve its quality.

To our knowledge there have not been attempts to define a generic incentive framework function that can be easily adapted to different crowdsourcing cases.

B. Industry and Crowdsourcing

From an industrial perspective, crowdsourcing has two main advantages. First, crowdsourcing delivers elasticity. Analogously to cloud computing, by working with the crowd, we have a virtually infinite number of resources that may be allocated and deallocated depending on the workload. Therefore, crowdsourcing offers flexibility in processes that include human beings. Secondly, it eliminates middleman costs. By building a platform to manage the crowd automatically, direct access to the final workers is gained, eliminating intermediate vendors that increase the cost of services.

As we mentioned in the introduction, the two main challenges regarding human interactions in a crowdsourcing platforms are: job quality evaluation and worker motivation.

Quality evaluation methods can be classified in three main families: (i) automatic, (ii) by direct inspection of the job provider and (iii) methods using the crowd itself as evaluator. Clearly for most of the tasks, an automatic evaluation is either impossible or can only guarantee a minimum quality, otherwise it would be possible to set up a completely automated solution without human intervention. Evaluation by the job provider has an inherent scalability problem, since the crowd can produce a large amount of work and the job provider has a finite amount of expert resources to evaluate it. The only way in which both problems can be overcome is by using the crowd for the evaluation, but this solution has the potential problem of trustworthiness and management of opinion and criteria disparity.

In terms of worker motivation, existing crowdsourcing platforms use a combination of the following extrinsic incentives to keep a working community engaged: (i) economical rewards, (ii) gamification, *e.g.*, public scoreboards, and (iii) free training. For instance, Mechanical Turk pays workers and Duolingo[3] gives free foreign language training while users actually translate strings, using gamification strategies to motivate users to improve their language skills (and thus translating more).

In [13] the authors propose an approach to deal with these two challenges (quality evaluation and worker motivation), by using a general mechanism to also crowdsource the quality evaluation of a job performed by the crowd and

[2] CA Technologies is a worldwide software and solutions provider that helps customers to make ICT management more agile, secure and flexible.

[3] http://www.duolingo.com/

give workers economical rewards based on the quality of their work. To build a trustworthy crowdsourcing system effectively, two essential aspects have to be addressed: mechanisms for worker coordination to guarantee the correct evaluation of quality, and a reliable mechanism to monitor the skills of workers. Is in this latter aspect that aggregation functions play a central role.

1) Crowd-based quality evaluation.: The general mechanism of [13] that guarantees job quality is based on the idea that human beings are the best quality evaluation method in many situations [5], [10]. Any complex task is subdivided into a series of subtasks called *Action-Verification Units* (AV-Unit). AV-Units establish relationship patterns between the workers of the crowd to help them to provide a higher degree of quality working in a collaboratively manner.

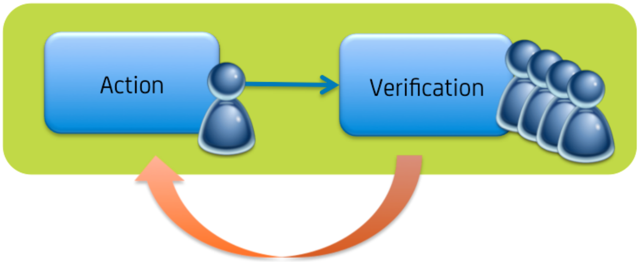

Fig. 1. Action-Verification Unit (AV-Unit) [13]

Figure 1 depicts an AV-Unit. As we observe, an AV-Unit is divided into two phases: Action and Verification. In the *Action* phase a single worker performs a specific action. In the *Verification* phase a set of workers verify the quality of the output generated in the previous Action phase. If the workers in the Verification phase consider that the quality provided is below a certain threshold, they might ask the first worker to repeat or improve the action. This process may be repeated iteratively until the output has reached a certain level of quality or the workers in the Verification phase decide to substitute the initial worker (or the worker is not available anymore). In practice, the Verification phase in the AV-Unit acts as a quality filter barrier, that does not allow to proceed with the process until the quality of each step in this process is approved by a set of human evaluators working collaboratively. Note that when more than one worker participates in the Verification phase, an aggregation function is also used to aggregate the decisions (scores) of each individual worker and produce a final decision (*i.e.*, the job has the required quality or not).

C. Aggregation Functions

Aggregation functions [21] are numerical functions used for information fusion that combine N numerical values into a single one. These operators are formally described as follows:

Definition 1: Let $X := \{x_1, \ldots, x_N\}$ be a set of information sources, and let $f(x_i)$ be a function that models the value supplied by the i-th information source x_i (for the sake of simplicity we often denote $f(x_i)$ by a_i), then a function $\mathbb{C} : \mathbb{R}^N \to \mathbb{R}$ is said to be an aggregation function if it satisfies:

1) $\mathbb{C}(a, \ldots, a) = a$ (unanimity, also known as idempotency)
2) $\mathbb{C}(a_1, \ldots, a_N) \leq \mathbb{C}(a'_1, \ldots, a'_N)$ if $a_i < a'_i$ (monotonicity)

There are several aggregation functions in the literature (see *e.g.* [4], [21] for further review). Among them, the most well-known aggregation functions are the arithmetic mean (AM) and the weighted mean (WM).

Yager defined in [24] the Ordered Weighted Averaging (OWA) aggregation function as a weighted linear combination of order statistics. In short, it works like a weighted mean after ordering the values a_i. We provide below a definition of the OWA operator using a non-decreasing function, as this is the most useful approach in our context.

Definition 2: Let Q be a non-decreasing function in $[0, 1]$ which satisfies the boundary condition $Q(0) = 0$ and $Q(1) = 1$, then the mapping $\text{OWA}_Q : \mathbb{R}^N \to \mathbb{R}$ defined as follows is an OWA operator:

$$\text{OWA}_Q(a_1, \ldots, a_N) = \sum_{i=1}^{N} \big(Q(i/N) - Q((i-1)/N)\big) a_{\sigma(i)}$$

where σ is a permutation of $\{1, \ldots, N\}$ such that $a_{\sigma(i)} \geq a_{\sigma(i+1)}$.

This operator has several properties. We underline the following ones:

i) For all Q, it holds that:

$$\min_i a_i \leq \text{OWA}_Q(a_1, \ldots, a_N) \leq \max_i a_i.$$

The choice of the function Q allows us to modulate OWA_Q from the minimum to maximum function. For example, when we consider the family of functions $Q_\alpha(x) = x^\alpha$, also called Yager Quantifiers, we have that large positive values of α lead to an OWA near to the minimum and, on the contrary, values of α near to zero lead to an OWA near to the maximum. Besides, when $\mathbf{a} = (a_1, \ldots, a_N)$ is fixed, OWA_{Q_α} is non-decreasing with respect to α.

ii) The OWA operator is symmetric for all Q. That is, the order of the parameters is not relevant for the computation of the output.

OWA operators are generalised by Choquet integrals [23] with respect to fuzzy measures, a family of fuzzy integrals that can be used for information fusion. In short, given a function f that represents the information supplied by the sources in X, the Choquet integral of f represents an aggregated value of those in f. In such integrals, fuzzy measures play the role of weights in the weighted mean.

Recall that a fuzzy measure μ is a set function over X such that the two boundary conditions hold (*i.e.* $\mu(\emptyset) = 0$ and $\mu(X) = 1$) and $\mu(A) \leq \mu(B)$ for every two subsets $A \subseteq B$ of X. A fuzzy measure is symmetric if it only depends on the cardinality of the set.

Formally, the Choquet integral is defined as follows:

Definition 3: Let μ be a fuzzy measure on X; then, the Choquet integral of a function $f : X \to \mathbb{R}^+$ with respect to the fuzzy measure μ is defined by

$$(C)\int f \, \mathrm{d}\mu = \sum_{i=1}^{N} a_{\sigma(i)}[\mu(A_{\sigma(i)}) - \mu(A_{\sigma(i-1)})]$$

where σ is a permutation such that $a_{\sigma(i)} \geq a_{\sigma(i+1)}$ and $A_{\sigma(i)} = \{x_{\sigma(1)}, \ldots, x_{\sigma(i)}\}$.

The Choquet integral with respect to the fuzzy measure $\mu(A) = Q(|A|/N)$ is precisely the OWA$_Q$ operator. This equivalence shows that OWA weights do not depend on the information sources nor possible relations between them. On the other hand, such independence is not required in the definition of a fuzzy measure and further aggregation functions can be defined with the Choquet integral (see *e.g.* [21] for a definition of the Weighted Ordered Weighted Averaging (WOWA) operator).

When a symmetric fuzzy measure is used, the Choquet integral is symmetric as the OWA operator. This property also holds for other fuzzy integrals. In particular, it also holds for the Sugeno integral [23]. Formally, the Sugeno integral is defined as follows:

Definition 4: Let μ be a fuzzy measure on X; then, the Sugeno integral of a function $f : X \to [0, 1]$ with respect to the fuzzy measure μ is defined by

$$(S)\int f \, \mathrm{d}\mu = \bigvee_{i=1}^{N} (a_{s(i)} \wedge \mu(A_{s(i)}))$$

where $\vee$ stands for maximum, $\wedge$ stands for minimum, s is a permutation such that $a_{s(i)} \leq a_{s(i+1)}$ and $A_{s(i)} = \{x_{s(i)}, \ldots, x_{\sigma(N)}\}$.

In particular, we can choose the symmetric fuzzy measure $\mu(A) = Q(|A|/N)$ as in the Choquet integral, obtaining an equivalent to the to the OWMax operator defined by Yager in [25].

Definition 5: Let Q be a non-decreasing function in $[0, 1]$ such that $Q(0) = 0$ and $Q(1) = 1$, then the mapping $SI_Q : \mathbb{R}^N \to \mathbb{R}$ defined as follows is a Sugeno integral of a function $f : X \to [0, 1]$ with respect to the fuzzy measure $\mu(A) = Q(|A|/N)$:

$$SI_Q(a_i) = \bigvee_{i=1}^{N} (a_{s(i)} \wedge Q(i/N))$$

where s is a permutation such that $a_{s(i)} \geq a_{s(i+1)}$.

The twofold integral [14], [20] is a generalisation for both Choquet and Sugeno integrals. The twofold integral is a fuzzy integral that aggregates a function with respect to two fuzzy measures. The rationale of this generalisation is that the semantics of both measures are different. In particular, the measure in the Choquet integral is seen as a 'probabilistic flavour' measure, and the measure used in the Sugeno integral is seen as a 'fuzzy flavour' measure. We use μ_C to denote the measure that corresponds to the one in the Choquet integral, and μ_S for the one in the Sugeno integral.

Definition 6: Let μ_C and μ_S be two fuzzy measures on X, then the *twofold integral* of a function $f : X \to [0, 1]$ with respect to the fuzzy measures μ_S and μ_C, denoted $TI_{\mu_S,\mu_C}(f)$, is defined by:

$$\sum_{i=1}^{n} \Big(\big(\bigvee_{j=1}^{i} a_{s(j)} \wedge \mu_S(A_{s(j)}) \big) \big(\mu_C(A_{s(i)}) - \mu_C(A_{s(i+1)}) \big) \Big)$$

where s is a permutation such that $0 \leq a_{s(i)} \leq a_{s(i+1)} \leq 1$ and $A_{s(i)} = \{x_{s(i)}, \ldots, x_{s(n)}\}$.

III. Using Worker Ranking for Trustworthiness Measuring

The success of AV-Units highly depends on the workers' profile. Involving many workers with low skills in an AV-Unit, might have a negative impact on the final quality. In most of industrial processes, quality standards are high and trusting the individuals in the crowd and their capacity to carry out the different tasks assigned to them is essential. Because of this, the main concern of a crowdsourcing platform is to monitor workers in order to evaluate and update their skills based on the quality of their past tasks.

To cope with this requirement, we propose to use a ranking systems that dynamically modifies the worker skills. Specifically, there are several aspects that might influence such a ranking system:

- ***T*he worker quality is measured from the output of their past jobs**: it is necessary to establish rewarding and penalty measures that modify the ranking of the workers in the crowd. In general, the actions with a higher impact for workers are those performed in the Action phase of an AV-Unit. However, it would be also possible to modify the ranking of workers based on their activity when they are acting in a Verification phase.
- ***G*eneral behaviour of the workers in the crowd**: other aspects might influence the ranking, *i.e.* worker commitment, career, etc. For instance, a worker might click very fast in order to get solutions quickly and get an economical reward. Although, improper behaviour will lead with high probability to bad quality, taking into account behavioural patterns may help to multiply the positive or negative impact of actions in the corresponding worker's ranking and speeding up the detection of incorrect behaviour.

The main idea behind using ranking systems is that a worker with a higher rank will be more trustworthy than workers with lower ranks. Therefore, it is possible to set a fair payment system where workers quality modulates the economical reward, increasing in this way the workers motivation to deliver high quality outputs in their future tasks. Additionally, ranking systems might allow us to automatically determine the most suitable workers for a given task.

A. Automatic Worker Ranking Determination

As aforementioned, a natural way to determine the rank of crowd workers is by combining the quality values of their previous works. This can be achieved by using aggregation functions. Concretely, foreach Action unit of previous AV-Units, denoted by x_i, that has at least one Validation value a_i, it is possible to combine all these quality values a_i in a single aggregated number $\mathbb{C}(a_1, \ldots, a_n)$. Note that, for a worker, this aggregated value cannot be bigger than the maximum quality achieved for any of his previous works nor lesser than the worst.

The choice of the aggregation function will lead to different ways of promoting and demoting crowd workers. For instance, the use of the arithmetic mean (AM) equally considers all previous works, independently on whether a task has been recently delivered or not. In this case, a bad quality task delivered during the training period counts exactly the same as a more recent task when the worker skills are better. To overcome this issue, we can replace the AM by the weighed mean operator (WM), doing this it is possible to give more importance to recent tasks than older ones.

On the other hand, by using OWA operators it is possible to emphasise extreme quality values. For example, using the fuzzy quantifier $Q(x) = x^\alpha$ with α close to 0, we are giving more importance to the highest quality tasks of a certain worker, without taking into account when the task was done (near in the past or not). In this setting, we are assuming that future tasks will also have a high quality because the worker finalised in the past (at least one time) a high quality task.

On the contrary, by using OWA operators with the same fuzzy quantifier but with $\alpha = 2$, the bigger the growth of the quantifier is when x values are near to one. Therefore, the associated weight of the low quality tasks will be bigger than the high quality ones. In this setting, we are penalising a lot workers with some low quality tasks. In this model, workers will find more difficulties to improve their rank. Finally, we can reduce the relevance of possible outliers by choosing a fuzzy quantifier $Q(x)$ with slow growth near $x = 0$ and $x = 1$.

In certain industrial scenarios, it may be interesting to consider both dimensions: task quality and when the task was delivered. To do this, it is necessary to use the twofold integral. In this integral, it is possible to include two fuzzy quantifiers, one for delivering time Q_t and another for task quality Q_q.

IV. Crowd-based Text Translation: A Practical Example

For this example we use a crowdsourcing platform for software and text localization (similar to text translation [15], [27]) that is being developed by CA Technologies (CA), applying the aforementioned AV-Units for ensuring task quality. In such platform, crowd workers are divided into three disjoint categories: *newcomers*, *associates* and *seniors*. Figure 2 presents a description of each category. Newcomers are workers that use the platform to learn and improve their professional translation and post-edition (reviewing manually the out put of an automatic translation) skills. They do not receive any economical reward for their translations and quality in real translations never depends on their work, but they receive feedback from senior translators to improve their skills. The platform also offers them many training examples. The main task of associate workers is to post-edit texts and their economic reward mainly depends on the text length. Finally, senior workers are those that know CA quality standards for translations and have proven very high skills in translation and post-edition. Their main task is to verify the work done by associates and provide them feedback to improve the quality of their translations and, as a consequence, their rank to become seniors. Note that, the economical reward of a senior translator is higher than that of an associate translator. The platform also considers that an associate translator becomes a senior translator if the quality of translations is high compared to the quality obtained by senior translators.

A. TQI, a Quality Measure For Text Translation

Translation Quality Index (TQI) [16] measure is the standard way to evaluate the quality of a professional translation. To compute the TQI, it is necessary that an expert review the translated text to detect the errors and evaluate their severity. Therefore, TQI measure reflects the criterion of such expert. To help experts to decide the severity level of an error, CA Technologies provides the following categories:

Sev1 Linguistic issues that bring the most direct (critical) impact to end users, such as:

- unexpected functional results, which are different from the English product description/statement
- deterring subsequent operations from product execution
- causing product features to fail
- carrying negative legal, political, security, and financial consequences or cultural noncompliance

Sev2 Linguistic issues that bring indirect yet substantial impact, leading the end user to:

- difficulties in understanding functionalities and technologies invented and developed in CA products
- misleading or incorrect interpretation of concepts for CA products

Sev3 Linguistic issues that do not impact the end user operation, yet impact the end user experience such as:

- spending additional time trying to figure out the meaning of descriptions/statements
- inconsistencies when trying to read product materials
- incompatibilities with layout and formatting

Sev4 Linguistic issues with minimal impact to overall end user experience, such as:

- noticeable and tolerable minor linguistic flaws
- layout and formatting errors that are only visible by comparing to the English source

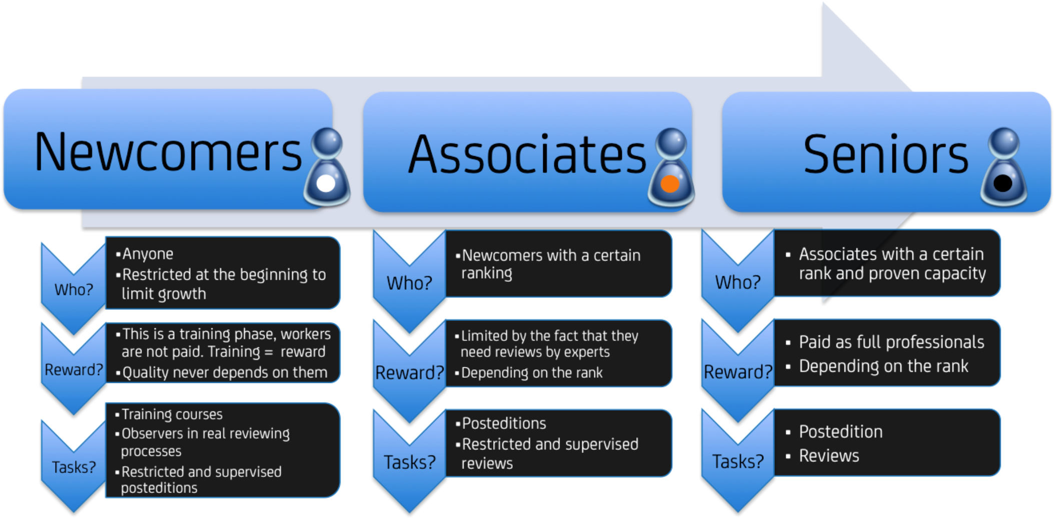
Newcomers
Associates
Seniors
Who?
•Anyone
•Restricted at the beginning to limit growth
Reward?
•This is a training phase, workers are not paid. Training = reward
•Quality never depends on them
Tasks?
•Training courses
•Observers in real reviewing processes
•Restricted and supervised postditions
Who?
•Newcomers with a certain ranking
Reward?
•Limited by the fact that they need reviews by experts
•Depending on the rank
Tasks?
•Postditions
•Restricted and supervised reviews
Who?
▪ Associates with a certain rank and proven capacity
Reward?
▪ Paid as full professionals
▪ Depending on the rank
Tasks?
▪ Postdition
▪ Reviews

Fig. 2. Translator categories at the CA Technologies crowdsourcing platform for localization.

TABLE I
TEXT QUALITY LEVELS BASED ON TQI RANKING.

Level	Criteria
Excellent	TQI $\geq$ 90
Good	80 $\leq$ TQI $<$ 90
Fair	70 $\leq$ TQI $<$ 80
Acceptable	60 $\leq$ TQI $<$ 70
Reject	TQI $<$ 60

Once all the translation errors have been detected and categorised, TQI is calculated as follows,

$$\text{Total} = (a_1 * 6) + (a_2 * 2) + (a_3 * 1.2) + (a_4 * 0.8)$$

$$\text{TQI} = 100 - (40 * (\text{Total}/\text{Reviewed Word}/0.01))$$

where a_i means the number of Sev i issues detected in the text. The final quality levels of a given translation are ranked in five categories, as it is depicted in Table I. In the next section, we use similar quality levels to decide whether a worker is ranked as newcomer, associate or senior.

B. AV-Units Applied to Text Translations

In this section, we present how to use AV-Units to create a crowd-based platform for text translation. The platform works as follows: Firstly, the source texts go through a rule-based machine translation engine and a first automatic translation is produced. Secondly, the original and the machine-translated texts are sent to external human translators who post-edit the text written in the target language, we call this step as post-edition step (PE), and it corresponds to the action part of the AV-Units methodology. Thirdly, multiple (from one to three) text verifications are performed by CA translators. It is important to note here that verifiers cannot modify text, they can only inform about detected errors. We call this part verification step (VE), and it corresponds to the verification part of the AV-Units. If the translation does not reach the minimum quality level (TQI $\geq$ 80), a new PE and VE steps are performed. To avoid infinite loops, in this second round verifiers are allowed to modify the text if they think it is required to maintain a TQI quality up to 80. After this process, the text is ready to be published.

C. Worker Categories and Promotion Mechanisms

As we have introduced before, workers are ranked into three different categories. To determine a worker category, we aggregate the TQI values obtained in the past into an overall TQI value, as it is explained in Subsection III-A. Table II depicts worker categories from the obtained overall TQI values. The platform assumes that the minimum overall TQI level for being considered an associate translator is equal to 60. Remember that translations with a TQI quality measure below 60 are rejected. For being considered a senior translator, the overall TQI obtained by a worker must be up to 80, since this is the minimum level required for being a CA internal translator.

In the platform, workers are automatically promoted or demoted depending on the overall TQI obtained in their past translations.

TABLE II
WORKER RANKING CATEGORIES.

Category	TQI value
Senior	TQI $\geq$ 80
Associate	60 $\leq$ TQI $<$ 80
Beginner	TQI $<$ 60

D. Numerical Examples

Experiments have been done using the weighted mean (WM), OWA operator and the twofold integral on three different associate workers (A, B and C). Table III shows their last 10 obtained TQI values (x_1 to x_{10}) together with

TABLE III

LAST TQI VALUES FOR WORKERS A, B AND C. x_{12} IS THE LAST ONE.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	$\mathbf{x_{11}}$	$\mathbf{x_{12}}$
A	65	63	62	64	66	71	72	74	77	78	**84**	**87**
B	72	65	64	70	68	69	73	72	70	72	**74**	**73**
C	77	80	76	74	75	70	68	65	61	60	**59**	**52**

TABLE IV

SUMMARY OF THE OBTAINED VALUES BY THE WEIGHTED MEAN (WM), THE OWA OPERATORS AND THE TWOFOLD INTEGRALS.

	WM		OWA$_{Q^e_{0.6}}$		OWA$_{Q^e_2}$		OWA$_{Q^s_{0.5}}$	
	10	12	10	12	10	12	10	12
A	74.2	80.1	71.6	75.6	66.0	67.4	68.7	71.1
B	70.9	71.9	70.5	71.2	67.9	68.5	70.0	70.7
C	64.6	59.6	73.2	71.3	66.9	63.4	71.4	68.8

	$TI(Q^t_{0.0})$		$TI(Q^t_{0.1})$		$TI(Q^t_{0.2})$		$TI(Q^t_{0.4})$	
	10	12	10	12	10	12	10	12
A	71.6	75.6	71.3	75.0	70.2	72.9	68.7	70.9
B	70.5	71.2	70.3	71.0	70.3	71.0	69.1	70.6
C	73.2	71.3	72.4	70.7	72.0	70.3	71.0	69.4

two new TQI values, in bold, (x_{11} and x_{12}) which must be considered to recalculate their new ranks. Worker A starts obtaining a bad performance, but at the end the quality of the translations are almost in the senior category. On the other hand, the quality of worker C is decreasing slowly. We want to test our aggregation functions by adding two more scores to each worker: A started to produce translations with a senior quality, worker B remains as an associate worker, and quality of worker C is on newcomer level. For the aggregation functions, three different families of fuzzy quantifiers were considered.

1) $Q^e_\alpha(x) = x^\alpha$
2) $Q^s_\alpha(x) = 1/(1 + e^{10(\alpha - x)})$
3) $Q^t_\alpha(x) = \begin{cases} 0 & \text{if } x \leq \alpha \\ 1 & \text{if } x > \alpha \end{cases}$

Here, Q^e stands for exponent, Q^s for sigmoidal and Q^t for threshold families. α values have been manually defined in each example.

In Table IV one can find a comparison between different aggregation functions before and after the addition of the two new TQI values (x_{11} and x_{12}). Weights $w_{N-i} = ((i+1)/N)^{0.3} - (i/N)^{0.3}$ in the WM were chosen assuming that the most recent TQI scores are the most relevant. Due to the extraordinary good performance of worker A in her last two translations, the WM operator promotes him to the senior category. On the other hand, worker C is penalised by his recent bad quality translations and his new category is newcomer. This model is appropriate for companies who want to frequently update their workers status.

By using OWA operators we focus on the TQI values themselves, instead of when they were obtained. For example, considering $Q^e_{0.6}$ we always reward the best translations ever done. In this case, worker C does not get much penalised because in his initial works he obtained very good TQI values. For the use of the OWA operator with the exponential

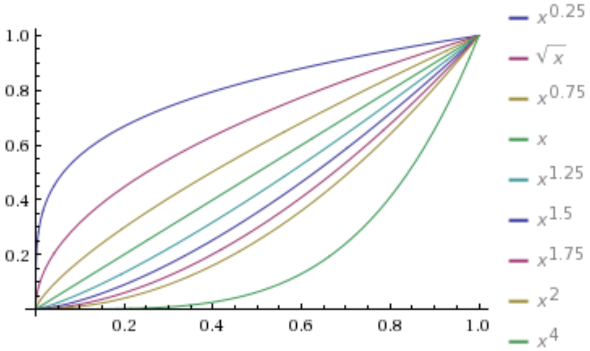

Fig. 3. Graphical representation of Q^e_α for several α values.

function Q^e_2, one can observe that the very good performance of worker A in his last tasks does not considerably affect his updated ranking. In this model, worker A needs to produce a large quantity of good quality translations in order to obtain the senior status. Finally, by using the sigmoidal function $Q^s_{0.5}$, we remove possible outliers in the TQI scores and focus on the average quality values. In this last example, the updated rank of the three workers is around 70 because this is their average task quality.

The twofold integral can be used as another tool to remove the relevance of possible outliers. The fuzzy measure $\mu_S(A) = Q^t_\alpha(|A|/N)$ will reduce the $100\alpha\%$ biggest TQI scores to a lower score obtained by the worker. For example, with $\alpha = 0.2$ we are reducing $a_{(1)}$ and $a_{(2)}$ in our example to $a_{(3)}$ and then we aggregate the values with an OWA operator. We can choose an operator that rewards good performance avoinding concerns about a potentially unfair fast growth in a worker score. In the underneath part of Table IV one can find a summary of the obtained values with the fuzzy measure $\mu_S(A) = Q^t_\alpha(|A|/N)$ for $\alpha = 0.0, 0.1, 0.2, 0.4$ and the fuzzy measure $\mu_C(A) = Q^e_{0.6}(|B|/N)$. In the $\alpha = 0.0$ case, the obtained values are exactly the same than in the $OWA_{Q^e_{0.6}}$ operator since we did not modified the original TQI scores. With $\alpha > 0.0$, one can see that the growth of worker A rank is not as large as in the original OWA operator.

V. CONCLUSIONS

In this paper we have studied the problem of combining past quality task evaluations using several aggregation functions to automatically determine worker category in a crowdsourcing platform holding a large professional community with different professional profiles. We have studied different cases to adjust the platform behaviour modifying the aggregation process (selected function and fuzzy measure). In this way we can set an automatic management system for worker promotions and demotions. We have also introduced a real crowd-based platform for text translations, describing how our ideas can be deployed in it.

In the near future, we will study the use of aggregation functions to determine the fairest economical reward for a given text, worker and final quality of the current translation.

References

[1] Luisa Bentivogli, Marcello Federico, Giovanni Moretti, and Michael Paul. Getting expert quality from the crowd for machine translation evaluation. In *MT Summit XIII*, pages 521–528, 2011.

[2] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. Soylent: a word processor with a crowd inside. In *Procs. of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 313–322, New York, NY, USA, 2010.

[3] D.C. Brabham. *Crowdsourcing*. MIT Press, 2013.

[4] Tomasa Calvo, G. Mayor, and Radko Mesiar. *Aggregation Operators*. Physica-Verlag, 2002.

[5] Qin Gao and Stephan Vogel. Consensus versus expertise: a case study of word alignment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 30–34, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[6] M. Hirth, T. Hossfeld, and P. Tran-Gia. Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms. In *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on*, pages 316–321, 2011.

[7] Jeff Howe. Wired 14.06: The Rise of Crowdsourcing, 2006.

[8] Ece Kamar and Eric Horvitz. Incentives for truthful reporting in crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '12, pages 1329–1330, Richland, SC, 2012. International Foundation for Autonomous Agents and Multiagent Systems.

[9] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2583–2586, New York, NY, USA, 2012. ACM.

[10] Dimitris Kontokostas, Amrapali Zaveri, S. Auer, and Jens Lehmann. Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data. In *Knowledge Engineering and the Semantic Web*, volume 394 of *Communications in Computer and Information Science*, pages 265–272, 2013.

[11] Afra J. Mashhadi and Licia Capra. Quality control for real-time ubiquitous crowdsourcing. In *Proceedings of the 2Nd International Workshop on Ubiquitous Crowdsouring*, UbiCrowd '11, pages 5–8, New York, NY, USA, 2011. ACM.

[12] Winter Mason and Duncan J. Watts. Financial incentives and the ”performance of crowds”. *SIGKDD Explor. Newsl.*, 11(2):100–108, May 2010.

[13] Victor Muntés-Mulero, Patricia Paladini, Jawad Manzoor, Andrea Gritti, Josep-Lluís Larriba-Pey, and Frederik Mijnhardt. Crowdsourcing for industrial problems. In *Citizen Sensor Networks*, volume 7685 of *Lecture Notes in Artificial Intelligence*, pages 6–18. Springer, 2013.

[14] Yasuo Narukawa and Vicenç Torra. Twofold integral and multi-step choquet integral. *Kybernetika*, 40(1):39–50, 2004.

[15] Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. Divide and conquer: crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 670–679, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[16] Riccardo Schiaffino and Franco Zearo. Developing and using a translation quality index. *Multilingual*, July/August 2006.

[17] T. Schulze, D. Nordheimer, and M. Schader. Worker perception of quality assurance mechanisms in crowdsourcing and human computation markets. In *Proc. of the 19th Americas Conf. on Information Systems*, 2013.

[18] Thimo Schulze, Dennis Nordheimer, and Martin Schader. Worker perception of quality assurance mechanisms in crowdsourcing and human computation markets. In *19th Americas Conference on Information Systems 2013 : AMCIS 2013 Proceedings*, Atlanta, Ga., 2013. AISeL.

[19] M. Six Silberman, Lilly Irani, and Joel Ross. Ethics and tactics of professional crowdwork. *XRDS*, 17(2):39–43, December 2010.

[20] Vicenç Torra. Twofold integral: A choquet integral and sugeno integral generalization. *Butlletí de l'Associació Catalana d'Intel·ligència Artificial*, 29:13–19 (in Catalan). Preliminary version: IIIA Research Report TR–2003–08 (in English)., 2003.

[21] Vicenç Torra and Yasuo Narukawa. *Modeling decisions: Information Fusion and Aggregation Operators*. Springer, 2007.

[22] Petros Venetis and Hector Garcia-Molina. Quality control for comparison microtasks. In *Proceedings of the First International Workshop on Crowdsourcing and Data Mining*, CrowdKDD '12, pages 15–21, New York, NY, USA, 2012. ACM.

[23] Michio Sugeno Vicenc Torra, Yasuo Narukawa. *Non-Additive Measures*. Springer, 2014.

[24] Ronald Yager. On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on System, Man, and Cybernetics*, 18:183–190, 1988.

[25] Ronald Yager. Applications and extensions of OWA aggregations. *International Journal Man-Machine Studies*, 37:103–122, 1992.

[26] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. A Survey of Crowdsourcing Systems. In *Proceedings of the IEEE Third International Conference on Social Computing (SocialCom)*, pages 766–773. IEEE, October 2011.

[27] Omar F. Zaidan and Chris Callison-Burch. Crowdsourcing Translation: Professional Quality from Non-Professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, USA, June 2011.

[28] Shaojian Zhu, ShaunKane, Jinjuan Feng, and Andrew Sears. A crowdsourcing quality control model for tasks distributed in parallel. In *Extended Abstracts on Human Factors in Computing Systems*, pages 2501–2506, 2012.