8-2010

Validating results of human-based electronic services leveraging multiple reviewers

Robert Kern
Karlsruhe Institute of Technology (KIT), robert.kern@kit.edu

Cordula Bauer
Karlsruhe Institute of Technology (KIT), cordula.bauer@gmx.net

Hans Thies
Karlsruhe Institute of Technology (KIT), hans.thies@gmx.de

Gerhard Satzger
Karlsruhe Institute of Technology (KIT), gerhard.satzger@kit.edu

Validating results of human-based electronic services leveraging multiple reviewers

Robert Kern
Karlsruhe Institute of Technology (KIT)
robert.kern@kit.edu

Cordula Bauer
Karlsruhe Institute of Technology (KIT)
cordula.bauer@gmx.net

Hans Thies
Karlsruhe Institute of Technology (KIT)
hans.thies@gmx.de

Gerhard Satzger
Karlsruhe Institute of Technology (KIT)
gerhard.satzger@kit.edu

ABSTRACT

Crowdsourcing in the form of human-based electronic services (people services) provides a powerful way of outsourcing so called micro tasks to large groups of people over the Internet in order to increase the scalability and productivity of business processes. However, quality management of the work results continues to be a challenge. Most existing approaches assume that multiple redundant results delivered by different people for the same task can be aggregated in order to achieve a reliable result, but for a lot of task types an automatic aggregation or comparison of task results is not possible. Also, cost considerations and estimators for outgoing quality have experienced little attention. Our *majority review* approach addresses these challenges by leveraging the crowd not only for delivering work results but also for validating the results delivered by others. An evaluation in a business context confirms that the approach is capable of gaining reliable results.

Keywords

Crowdsourcing, micro task, quality, human-based, people service, review

INTRODUCTION

The idea of human-based electronic service is that they look like Web services but they are not performed by a computer, instead they use human workforce out of a crowd of Internet users. The success of Amazon's Mechanical Turk (MTurk) platform[1] and the growing number of companies that build their business model entirely on that platform demonstrate the potential of this approach. The MTurk platform acts as a broker between requesters who publish micro tasks and workers who perform those tasks in return for a small amount of money.

Kern et al. proposed the term *people services* (pServices) for this type of human-based electronic services and define it as "Web based software services that deliver human intelligence, perception, or action to customers as massively scalable resources" (Kern, Zirpins and Agarwal, 2009). As there is limited control over the individual contributors, particular attention has to be paid to the quality of the work results. Kern et al. argue that platform support is needed for pServices in order to guarantee a certain level of quality. They propose a platform that brings together *service requesters* who submit pService requests (tasks) and *service workers* who do not only work on those requests but are also leveraged for ensuring the quality of the results using coordination mechanisms like redundant execution or review processes.

The aim of this paper is to propose and evaluate the *majority review* approach as one possible approach for ensuring the quality of pServices by leveraging the crowd of service workers for reviewing the results submitted by others.

After providing the theoretical foundation, the majority review approach is introduced and the maximum likelihood estimation is adapted to it as a possible way of weighting and aggregating the contributions of the individual reviewers. Thereafter, the approach is evaluated based on a business scenario of a pService platform provider. The paper concludes with a summary.

THEORETICAL FOUNDATION

Existing approaches for ensuring the quality of pService results

For annotation tasks, Sorokin and Forsyth (Sorokin and Forsyth, 2008) distinguish between two quality assurance strategies on pServices platforms which leverage the crowd itself for quality assurance: the collection of multiple annotations and the

[1] www.mturk.com

performing of a separate review task (they call it grading task). In order to establish a common terminology we propose the following terms and more general definitions for the two approaches:

- *The **majority vote** approach introduces redundancy by passing the same task to multiple workers and aggregating the results in order to compute the result with the highest probability for correctness.*
- *The r**eview** approach leverages individuals of the crowd for reviewing (e.g. validating) the results delivered by others.*

Majority vote mechanisms are already widely used in pServices scenarios today, e.g. on the MTurk platform. Initial research has shown that an amazing level of result quality can be achieved for basic tasks like natural language annotation, image labeling and data labeling: Sorokin & Forsyth (Sorokin and Forsyth, 2008) identify objects on images by combining the drawings of silhouettes of distinct persons. Snow et al. let ten distinct workers rate natural language expressions and compare their aggregated results to gold-standard labels given by experts (Snow, O'Connor, Jurafsky and Ng, 2008). They find a high degree of agreement between the individual workers and the gold-standard labels.

Barr et al. mention the application of review for quality control on MTurk (Barr and Cabrera, 2006). They state that it is common to use a secondary task to verify the result of the original task. As an example they mention translations were the original task is the actual translation and the second is the examination of the translation to verify its accuracy. They find that these two tasks require different skills. While for the first it is essential to know the two languages necessary for the translation, the second does not require bilingual abilities but a good command of the language the text should be translated to. Another example for a review technique on MTurk is the mechanism that CastingWords[2] uses for transcription of audio and video files. According to Hoffmann (Hoffmann, 2009), a transcript created by one worker is rated and improved in multiple steps by a series of other workers until the required level of quality is reached.

There are no scenarios known to the authors so far, in which a review mechanism is formally modeled and examined in the context of pServices. As a starting point, this paper focuses on developing a formal description of a review scenario which leverages multiple reviewers per task.

Determining inter-observer reliability using the maximum likelihood estimation

Dawid and Skene (Dawid and Skene, 1979) developed an approach to iteratively determine the most likely category for a subject when multiple observers can classify the subject into several categories. The method also derives error rates for the observers. The approach is based on the EM algorithm, which dates back to Dempster et al. who introduced it as a generalized method to compute maximum likelihood estimates from incomplete data (Dempster, Laird and Rubin, 1977). The basic idea of the algorithm is to infer the posterior probability of the true category of a subject by integrating its a priori probability, the observers' judgments and the observers' error rates via the Bayes Theorem. This is done iteratively until the posterior probabilities converge. The algorithm is also used in (Snow et al., 2008) to estimate the error rates of annotators in an MTurk experiment.

Intuitively with K observers k = 1,…K and J possible categories j=1,…J individual error rates $\pi_{jl}^{(k)}$ are defined as

$$\boldsymbol{\pi_{jl}^{(k)} = \frac{number\ of\ cases\ observer\ k\ records\ category\ l\ when\ j\ is\ correct}{number\ of\ cases\ assesed\ by\ observer\ k\ where\ category\ j\ is\ correct}} \quad (1)$$

If the true category is not known, the $\pi_{jl}^{(k)}$s must be estimated. Dawid and Skene's approach of this estimation along with the maximum likelihood estimation of the true categories of each subject, is described in the following.

Let I be the set of all subjects i=1,…I and let T_{ij} be the probability that subject i belongs to category j. If the true category of subject i is known, we would have the situation where $T_{iq} = 1$ for the true label q and $T_{ij} = 0$ for $j \neq q$. If the true category is not known the T_{ij} s take values between 0 and 1.

Let $n_{ij}^{(k)}$ be the number of times that observer k states that subject I belongs to category J. The maximum likelihood estimates of the observer error rates can now be assessed by

$$\hat{\pi}_{jl}^{(k)} = \frac{\sum_{i=0}^{I} T_{ij}}{\sum_{l=0}^{J} \sum_{i=0}^{I} T_{ij} \cdot n_{il}^{(k)}} \quad (2)$$

[2] http://castingwords.com

Note that (2) is semantically equal to (1).

The probabilities of the categories $p_j\,(j = 1, \ldots J)$ can be estimated via

$$\hat{p}_j = \sum_{i=0}^{I} \frac{T_{ij}}{|I|} \tag{3}$$

The challenge is now to find adequate values for $\{T_{ij}\}$. If $\{p_j\}$ and $\{\pi_{jl}^{(k)}\}$ would be known, Bayes' Theorem can be used to obtain estimates of $\{T_{ij}\}$. We would obtain

$$p(T_{ij} = 1|data) = p(data|T_{ij} = 1) \cdot p(T_{ij} = 1)$$

what in numerical terms is

$$p(T_{ij} = 1|data) = \frac{\prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{jl}^{(k)})^{n_{il}^{(k)}} \cdot p_j}{\sum_{q=1}^{J} \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{ql}^{(k)})^{n_{il}^{(k)}} \cdot p_q} \tag{4}$$

If the p's and the π's are not known but estimated, the calculation does still work and give as a maximum likelihood estimation of the T's. The point is to find a good initial estimate. Dawid & Skene propose the following iterative procedure:

1. Obtain initial estimates for the T's.
2. Use equations (2) and (3) to estimate the p's and the π's.
3. Use equation (4) and the estimates of the p's and the π's to get new estimates of the T's
4. Repeat steps (2) and (3) until the T's converge.

As starting values for the T's they propose a relation of the judgments $n_{ij}^{(k)}$:

$$\hat{T}_{ij} = \frac{\sum_{k=1}^{K} n_{ij}^{(k)}}{\sum_{k=1}^{K} \sum_{l=1}^{J} n_{il}^{(k)}} \tag{5}$$

THE MAJORITY REVIEW APPROACH

Scope and relevance

The majority review approach investigated in this paper belongs to the category of review approaches. Before describing the majority review approach in detail, this section aims to provide some general considerations about the relevance of the review mechanism compared to majority vote.

Regarding the majority vote approach, it is obvious that it cannot be used for all possible types of pService tasks. As it relies on the aggregation or comparison of task results delivered by multiple workers, it implicitly assumes that the results can be aggregated or compared. However, there are many types of tasks for which this is not the case. For example, if several workers are asked to write a small abstract about a specific movie or video, the resulting abstracts cannot be simply automatically aggregated or merged into a single aggregated result. A comparison of the various abstract won't work either because there are simply too many possibilities of writing an abstract about a movie. In other words, the result space for the task is too large.

We propose the term *deterministic task* for such tasks, for which there is a well defined optimal result i.e. for which two workers who perfectly meet the task objectives will pass exactly the same results, or for which the responses can at least be automatically transformed (normalized) into a well defined optimal result. Even translation tasks in which workers are asked to translate a text into a foreign language are *non-deterministic* because there are many different correct translations for a given text.

From that point of view, majority vote approach is not appropriate for non-deterministic tasks. It can be applied neither to the authoring nor to the translation scenario described above. Of course, for both examples, semantic algorithms could support an aggregation of the abstracts or of the translations at least to some extent. This, however, would introduce additional task specific complexity which would widen the challenge beyond the scope of the majority vote approach. Without such additional complexity the majority vote approach cannot be applied to non-deterministic tasks.

The review approach, however, can obviously even be used for ensuring the quality of non-deterministic results because a human reviewer has the flexibility to deal with variety. This fact underlines the relevance of the review approach and implicitly of the majority review approach discussed in the following sections.

In addition to the extended reach of the review approach, there is a series of other advantages compared to majority vote. Depending on the task type it can be much more cost efficient because reviewing a result can be much faster than generating it. Furthermore, as described in the result section, the review scenario provides a powerful way of passing feedback to workers which can help them to improve their skills.

Basic scenario

The majority review approach is a combination of the two approaches that we referred to as *majority vote* and *review* above. Two or more service workers rate the same task result provided by another worker and their judgments are aggregated in order to compute the most likely overall rating. In the simplest case which is examined in this paper, the rating is a binary decision whether to accept or reject the result. Figure 1 shows a schematic description of the majority review scenario. After a *raw result* has been returned by a worker (step 1), a series of *review tasks* are generated (step 2) which are completed by a number of *reviewers* and aggregated into a single *consolidated review* (step 3). This consolidated review is used in two ways:

1. For rating the original *raw result* and either accept it or reject it (step 4) which leads to a *validated result* or a *rejected result* (step 5).
2. For evaluating the contributions of the individual reviewers by comparing their *raw reviews* with the *consolidated review* and updating the individual error rates accordingly (step 6).

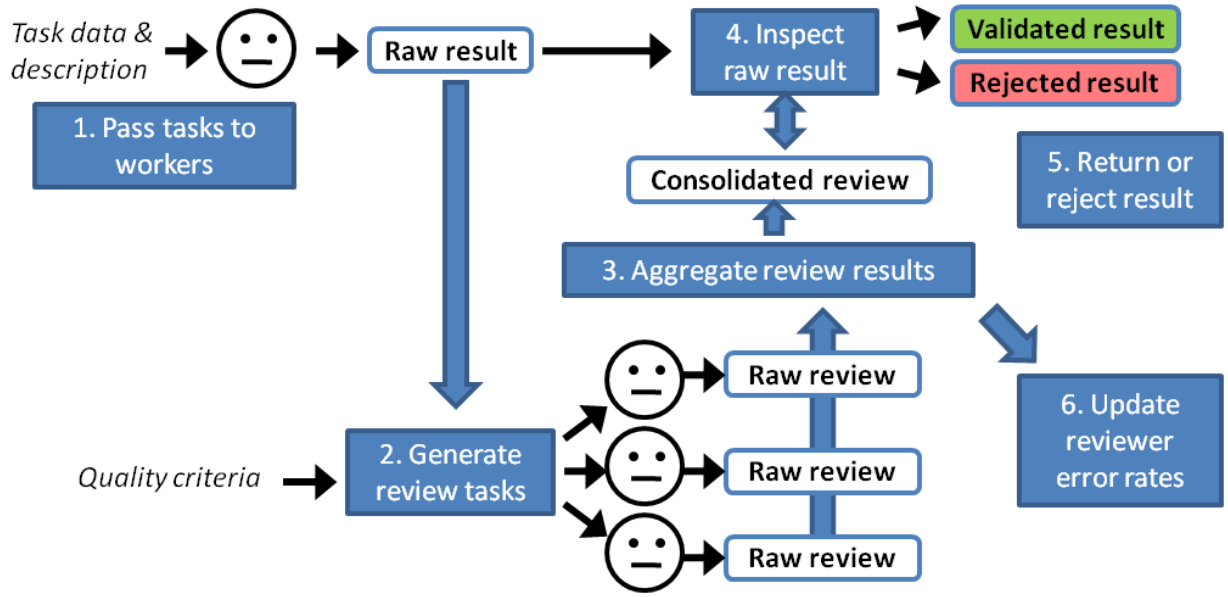

Figure 1: Schematic description of the majority review approach

To summarize, the scenario consists of the following entities:

- *Tasks* comprise a task description, task data, and *quality criteria*.
- *Raw results* can be either correct or incorrect according to the quality criteria of a task.
- *Raw reviews* represent a binary decision of a reviewer whether to accept or reject a raw result.
- *Workers* turn tasks into raw results and have an individual error rate p when working on tasks.
- *Reviewers* validate raw results and return raw reviews. They have individual error rates e1 (type I) and e2 (type II) when validating results submitted by other workers.

Formal model

In a more general form of the majority review scenario described in Figure 1, the steps 3 to 6 do not necessarily have to be performed separately for each raw result that is to be reviewed. In case the scenario does not require a validation to be

performed immediately, a more profound calculation of the consolidated reviews can be performed by evaluating a batch of multiple raw results at once. This is exactly what will be done by applying the maximum likelihood method which was introduced on page 2:

The following representation of the scenario assumes that there is a sequence N_t of t batches of raw results i, each being reviewed by k_t reviewers r who return the judgment $G_i^{(r)}$about the correctness of the raw result i. $G_i^{(r)}$ is either 0 if the reviewer states that the task is conforming or 1 if he states that it was non-conforming.

When leveraging reviewers out of the crowd and taking the service requesters' quality understanding as a benchmark it is obvious that the reviewers perform errors. Two types of errors can occur: we speak of a type I error if the service worker denotes a conforming task as nonconforming. Respectively a type II error occurs if the service worker classifies a task as nonconforming when it is actually conforming. A reviewer's type I error for batch t will be denoted as $e_{1,t}^{(r)}$ and a type II error respectively as $e_{2,t}^{(r)}$. The challenge is to estimate the error rates $e_{1,t}^{(r)}$ and $e_{2,t}^{(r)}$ as well as the task quality T_i and the fraction of non-conforming tasks p_t in each batch.

The given parameters are:

- N_t is the batch of raw results i to be rated per time interval
- k_t is the number of reviews per task and per time interval
- R is a set of reviewers
- $G_i^{(r)}$ is the judgment of reviewer r regarding the quality of task i. As not all reviewers are involved in the reviews for all tasks, $G_i^{(r)}$ is only defined for $i \in K_i$ with

$$K_r = \{i \mid reviewer\ r\ gave\ a\ judgment\ on\ the\ result\ of\ task\ i\}$$

The following parameters are to be determined by leveraging the maximum likelihood method:

- $e_{1,t}^{(r)}$ is the estimated type I error rate of reviewer r per time interval t
- $e_{2,t}^{(r)}$ is the estimated type II error rate of reviewer r per time interval t
- T_i is the estimate for the true result rating of task i
- p_t is the estimated fraction nonconforming per time interval t

Applying the maximum likelihood method

The challenge to be addressed by the maximum likelihood method is to find estimates for these errors and for the true fraction nonconforming of the task results which are good enough to meet the requirements of the service requester.

When applying Dawid and Skene's approach to expectation maximization, we make an important assumption: we presume that the majority of the reviewers know the true quality of the task. This is very important for the estimation to work at all.

The K observer in Dawid and Skene's approach correspond to our R reviewers. The J categories correspond to our two possible judgments over the quality of results: conformance and nonconformance. The set of subjects I correspond to the set N_t of raw results to be rated per time interval. The number of times observer k states that subject i belongs to category j, namely $n_{ij}^{(k)}$, corresponds to judgments of the R reviewers $G_{ij}^{(r)}$. A difference here is that in our model the values are binary and not enumerative because a reviewer gives only one judgment per task and not various. Thus, if reviewer r thinks that task i belongs to category j, then we obtain $G_{ij}^{(r)} = 1$, otherwise $G_{ij}^{(r)} = 0$. As we only deal with two categories, we have $G_{i0}^{(r)}$ and $G_{i1}^{(r)}$ that always add up to one, because one of the categories has to be chosen. Therefore, we define only $G_i^{(r)} = G_{i1}^{(r)}$. As not all reviewers judge each task, $G_i^{(r)}$ is only defined of if reviewer r gives a judgment about the quality of task i.

It remains to transfer the concept of the probability that subject i belongs to class j which was denoted as T_{ij}. In our case T_{i1} would be the probability the task result is conforming, $p(Q_i = 1)$, and T_{i0} the probability that it is not, $p(Q_i = 0)$. Since T_{i1}= 1- T_{i0}, we actually only need one value and thus we define $T_i = T_{i1}$.

The most likely fraction nonconforming is the estimated probability of category "nonconforming" (j = 1) per time interval t. It can be estimated via

$$\hat{p}_t = \sum_{i \in N_t} \frac{T_i}{|N_t|}$$

The probability of category "conformant" is just $1 - \hat{p}_t = \sum_{i \in N_t} \frac{(1-T_i)}{|N_t|}$. For reasons of simplicity, we will omit the index t in the following.

For all reviewers that have been involved in at least one review, the type I error is calculated with

$$e_1^{(r)} = \pi_{01}^{(r)} = \frac{\sum_{i \in K_r} (1 - T_i) \cdot G_i^{(r)}}{\sum_{i \in K_r} (1 - T_i) \cdot (1 - G_i^{(r)}) + \sum_{i \in K_r} (1 - T_i) \cdot G_i^{(r)}}$$

Respectively r's type II error rate is obtained via

$$e_2^{(r)} = \pi_{10}^{(r)} = \frac{\sum_{i \in K_r} T_i \cdot (1 - G_i^{(r)})}{\sum_{i \in K_r} T_i \cdot (1 - G_i^{(r)}) + \sum_{i \in K_r} T_i \cdot G_i^{(r)}}$$

The calculation of $\pi_{00}^{(r)}$ and $\pi_{11}^{(r)}$ as reviewer r's success rates is done analogously.

The calculation of the posterior probabilities of an error in solution to task i is

$$p(T_i = 1|data) = p(data|T_i = 1) \cdot p(T_i = 1)$$

what in numerical terms is

$$p(T_i = 1|data) = \frac{\prod_{r \in R} \left(\pi_{10}^{(r)}\right)^{1-G_i^{(r)}} \cdot \left(\pi_{11}^{(r)}\right)^{G_i^{(r)}} \cdot p}{\prod_{r \in R} \left(\pi_{00}^{(r)}\right)^{1-G_i^{(r)}} \cdot \left(\pi_{01}^{(r)}\right)^{G_i^{(r)}} \cdot (1-p) + \prod_{r \in R} \left(\pi_{10}^{(r)}\right)^{1-G_i^{(r)}} \cdot \left(\pi_{11}^{(r)}\right)^{G_i^{(r)}} \cdot p}$$

The iterative procedure described by Dawid & Skene remains unchanged.

Choosing appropriate starting values for the task quality estimation

As starting values for the T_i's we refer to the judgments $G_i^{(r)}$:

$$\hat{T}_i = \frac{\sum_{i \in K_r} G_i^{(r)}}{k}$$

If that initial estimation of the tasks' quality T_i is far away from their true quality, the algorithm will not deliver a satisfactory result. Therefore, error rates which have been determined in previous runs or during training phases should be leveraged to adapt the initial estimation of the T_i's.

EVALUATION

Background

We evaluated the majority review approach in a business scenario with kind support of the German pService platform provider bitworxx[3]. They state to have a workforce of over 10,000 workers available around the globe. A majority of their workforce are call center agents who work on the micro tasks in their idle time.

Test scenario

In our test scenario, the service requester is an online shopping platform on which traders can offer articles with minimal effort. The traders only need to specify the GTIN (*Global Trade Item Number*, formally *European Article Number* EAN-13) of the articles they want to sell. GTIN is the 13-digit barcode found on each product sold in the EU. It is a standard supported by *Global Standards One*[4]. While the GS1 maintains a catalogue with company prefixes[5], there is no central directory for the

[3] www.bitworxx.com

[4] http://www.gs1.org

[5] http://gepir.gs1.org/V31/xx

articles because the companies are responsible for allocating the remaining digits. For the shopping platform this implies that the article details have to be researched based on the EAN in order to provide a product description to the potential buyers. This task was outsourced to bitworxx.

The graphical user interface of the GTIN research task includes radio buttons to select a product category and product specific form fields to enter the article details including name and description. Furthermore, it contains a check box to indicate that no information about the GTIN was found.

Three types of errors are differentiated in the scenario: *no information found*, *non-conforming product category selected*, and *incorrect product details specified*. The service requester accepts an error rate of 10-15%, combining all three error types.

Gold standard

In order to ensure the result quality, an employee of the bitworxx team reviews all results returned by the workers. In case of incorrect results, individual feedback is passed to the workers to help them getting better over time. The review is also regarded as a way to communicate the requirements to the workers. Despite the high review effort, bitworxx has made very positive experience with this procedure, as the customers' quality requirements have always been met. Because of that, it was used as the baseline and reference for our tests i.e. the result of the bitworxx review is considered to be the gold standard which represents the true task result quality. The evaluation results will show that this point of view is somewhat problematic.

Test execution

The data for the experiment was collected within the real GTIN project. Between 20/01/2010 and 28/01/2010, a batch of 2002 tasks was processed as usual with each task handled by one worker and reviewed by the bitworxx team. For our test, six additional review tasks were created per task, which were processed on 7 working days. Ten people who were already experienced with the GTIN scenario supported the experiment, nine of them acted as workers and eight as reviewers i.e. most of them took both roles. They did neither know that they were supporting an experiment nor that the same review tasks were given to multiple workers.

Results

Worker and reviewer statistics

The left diagram of Figure 2 shows the amount of task results submitted per day and per worker, the area of the circles represents the number of tasks. The right diagram shows the number of data research tasks processed per worker as dark (blue) bars and the number of task results reviewed per reviewer as light-colored (red) bars.

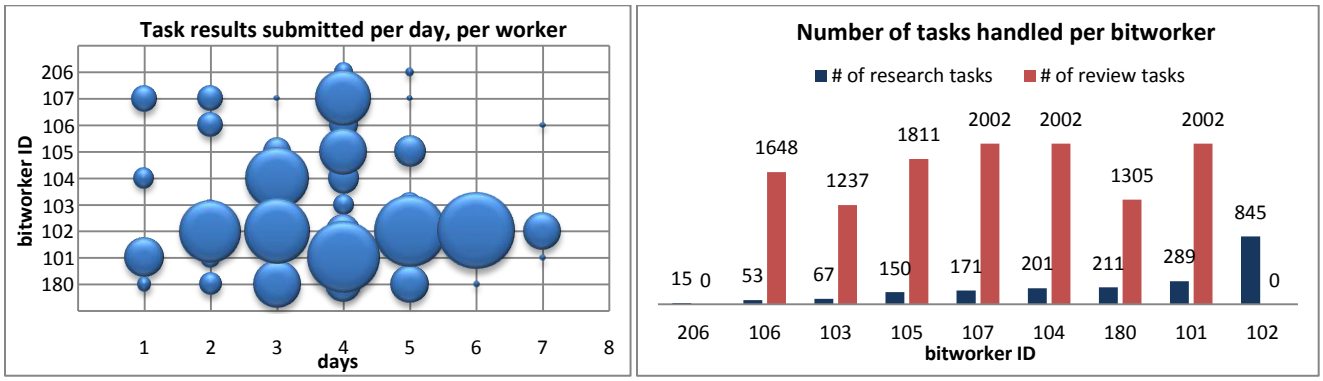

Figure 2: Results submitted per day, per worker (left), number of tasks handled per worker (right)

Estimations of the EM algorithm

The EM algorithm estimates not only the true result quality T_i but also the fraction of nonconforming results p_t as well as the type I and II error rates for all reviewers $e_{1,t}^{(r)}$ and $e_{2,t}^{(r)}$. Figure 3 compares these values with the actual values determined based on the gold standard: The left chart shows the estimated fraction nonconforming $p\ (EM)$ returned by the EM algorithm with the actual fraction nonconforming $p\ (gold)$ determined by the gold standard. The right chart compares the weighted average of the reviewer error rates e1(EM) and e2(EM) with the weighted average of the actual reviewer error rates e1(gold) and e2(gold). The weighted average takes the different number of tasks into account that a worker or reviewer has worked on.

Obviously, the EM algorithm overestimates the fraction nonconforming, and the type I error rate constantly, especially on day one. The deviations from the type II error rate are alternating.

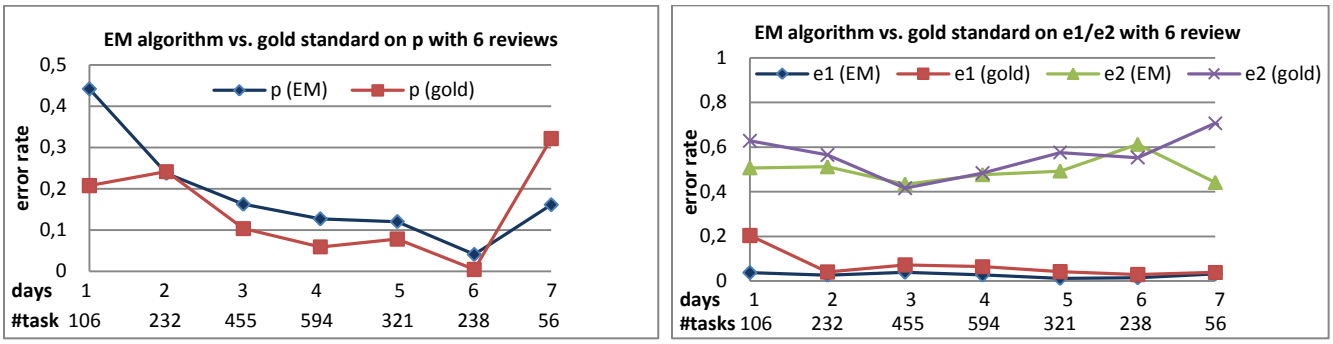

Figure 3: Estimations made by the EM algorithm vs. gold standard

Result ratings of the EM algorithm

From the perspective of the majority review approach, the primary output of the EM algorithm are the estimates for the true result rating T_i. According to the model, these estimates finally decide whether a result returned by a worker is accepted or rejected. The gold standard was used to determine to what extent these estimates can be regarded correct. A type I error happens if the EM algorithm rates the result nonconforming although the gold standard claims it to be conforming. A type II error happens in the opposite case. Figure 4 shows the correctness of the ratings for each error type, type I (left) and type II (right). The average type I error rate is 0.138 while the average type II error rate is 0.398.

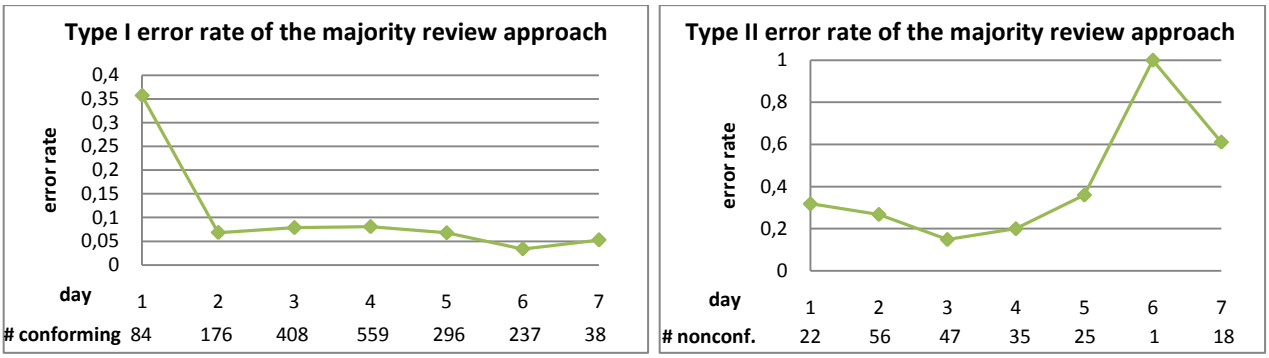

Figure 4: Actual type I (left) and type II (right) error rate of the majority review approach

Discussion

Although the error rates of the majority review approach are high in our scenario, the approach turns out to be capable of meeting the customer requirement of a maximum of 10 percent incorrect results: With a type II error rate of 0,398 some 60 percent of the 204 nonconforming results are detected which leads to 123 nonconforming (6.14 %) results being returned to the service requester. This assumes that all results being identified as nonconforming by the majority review approach are being successfully reworked by the workers.

The left chart of Figure 4 shows a relatively high rate of type I errors at the first day. We see two possible explanations for the phenomena: A first explanation could be that the reviewers were extraordinarily strict on the first day, because it was their first day as reviewers at all. A second explanation would be that the bitworxx reviewer was careless on the first day for some reason. And in fact, a series of cases has been identified in which the bitworxx reviewer did obviously not perform the task well.

Figure 5 gives an example. The table presents judgments of the reviewers and the bitworxx team together with comments for a concrete task. The worker who completed the task had claimed that he did not have found information about the EAN 21165105287. While the bitworxx reviewer agreed with the worker, four out of six reviewers found information on the Web and some even posted the corresponding links. By checking them, the requested product information can in fact be found, which confirms that the task result is non-conforming. What was suspected above becomes obvious here: The review performed by bitworxx is not a real gold standard and thus all comparisons with it have to be regarded with care.

The type II error rates shown on the right chart of Figure 4 are even higher than type I error rates, even though one should not be irritated by the outlier on day 6, which was based on only one nonconforming task. Both the type I and even more the type II error rate can be a sign for the careless handling of the review task by the reviewers and / or the unclear communication of the service requester quality requirements.

Reviewer	Rating	Comment
bitworxx	conforming	
101	non-conforming	##1##http://www99.shopping.com/xPO-Saitek-Saitek-Mini-Color-UFO-Hub-Metallic-Blue##1##
107	non-conforming	##1##http://www1.shopping.com/xPO-Saitek-Saitek-Mini-Color-UFO-Hub-Metallic-Blue##1##
104	non-conforming	##1## Please also use Google for data research… ##1##
105	conforming	
180	non-conforming	##1##http://www.shopping.com/xPO-Saitek-Saitek-Mini-Color-UFO-Hub-Metallic-Blue ##1##
106	non-conforming	##1## Infos found with Google ##1##

Figure 5: Judgments for job with id=77176 (translated from German)

In order to examine the variations between the ratings of the EM algorithm and the actual result quality defined by the gold standard, the agreement between the reviewers and the bitworxx team was examined on a task level. Figure 6 plots the frequency of a certain number of judgments on nonconformance for actually conforming results (left chart) and the frequency of a certain number of judgments on conformance for actually nonconforming results (right chart).

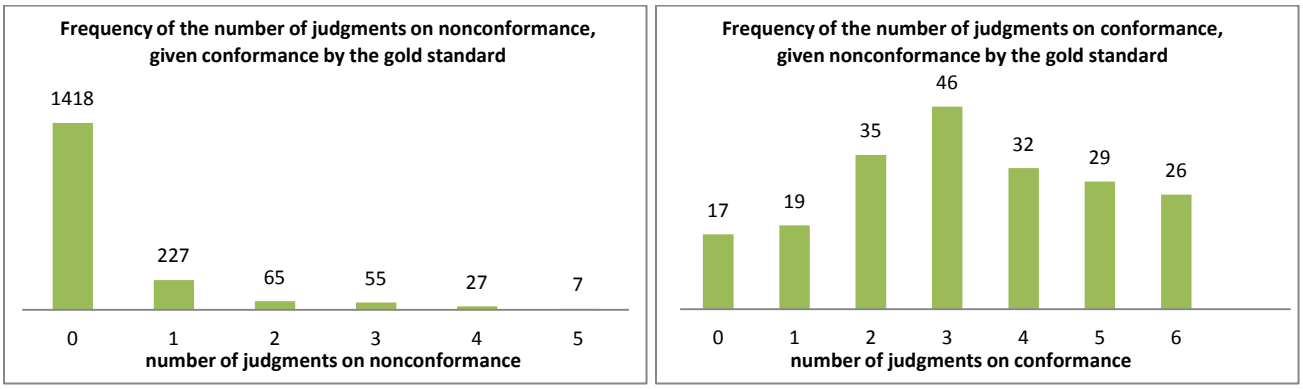

Figure 6: Frequency of number of judgments on nonconformance (conformance), given conformance (non-conformance)

On the left chart, for 1418 tasks all reviewers came to the same conclusion as the bitworxx review and classified the task result as conforming. However, for 154 tasks (8.5%) at least two reviewers classified the task as nonconforming, while it was classified as conforming by the bitworxx reviewer. This could be the reason for the overestimation of the fraction nonconforming.

The right chart shows that there is a high disagreement among the reviewers when the task is classified as nonconforming by bitworxx. In 42.6% of the cases the majority of the reviewers declared the task result as conformant when the bitworxx review did not. On the one hand, this might be a sign that the service requesters' requirements were not communicated satisfactorily and on the other hand it might show that the review task was often carelessly handled by the reviewers. Fortunately, even for the cases where only a minority claimed the task to be nonconforming, the EM algorithm classified it as nonconforming. This is because the algorithm recognizes the individual type II error rates and takes them into account iteratively when calculating the most likely overall review rating.

For a series of tasks, only the bitworxx reviewer rejected a result while all other reviewers accepted the result. Possibly only the bitworxx reviewer reflected the service requesters' perception because the work instruction had not been specified precisely. It might also be an example that the majority can come to a better conclusion than a single expert. This is exactly

what Surowiecki is arguing for in his book "The Wisdom of Crowds" (Surowiecki, 2004) and the great benefit of pServices is that they are able to profit from this wisdom.

CONCLUSION

We have proposed a *majority review* approach that can be used to ensure the quality of human-based electronic services (people services) by leveraging a group of reviewers for validating task results. The so called EM algorithm was applied to the approach in order to calculate the maximum likelihood estimation for the consolidated rating of the reviewers along with the estimated reviewer error rates and the fraction nonconforming. An evaluation within a business scenario of a people service provider has confirmed that the majority review approach can be successfully used for verification of task results. The results underline the importance of a clear definition of quality requirements but they also show that anonymous workers in the internet can come to better solutions than a single expert. The results also suggest that reviewers tend to rubber-stamp the tasks, which leads to high type II error rates. Fortunately, the EM algorithm identifies and omits these errors when calculating the most likely task result.

In our ongoing research we are extending our approach by concepts out of the field of statistical quality control (SQC) in order to be able to reach a predefined level of result quality while further reducing the overall quality management effort.

REFERENCES

1. Barr, J. and Cabrera, L.F. (2006) AI gets a brain, *ACM Queue*, 4(4), 24-29.

2. Dawid, A. and Skene, A. (1979) Maximum likelihood Estimation of Observer Error-Rates using the EM algorithm, *Journal of the Royal Statistical Society*, 28(1), 20-28.

3. Dempster, A., Laird, N. and Rubin, D. (1977) Maximum Likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, 39(1), 1-38.

4. Hoffmann, L. (2009) Crowd Control, *Communications of the ACM*, 52(3), 16-17.

5. Kern, R., Zirpins, C. and Agarwal, S. (2009) Managing Quality of Human-Based eServices, in *ICSOC '08: Proceedings of the International Conference on Service-Oriented Computing - Workshops*, Berlin, Germany (et al.): Springer-Verlag, pp. 304-309.

6. Snow, R., O'Connor, B., Jurafsky, D. and Ng, A. (2008) Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, USA: ACL, pp. 254-263.

7. Sorokin, A. and Forsyth, D. (2008) Utility data annotation with Amazon Mechanical Turk, in *CVPRW '08: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*, Washington, USA: IEEE Computer Society, pp. 1-8.

8. Surowiecki, J. (2004) *The Wisdom of Crowds*, Doubleday, New York, USA.