Quality Control of Crowdsourcing through Workers Experience

Li Tai¹, Zhang Chuang¹, Xia Tao¹, Wu Ming¹ and Xie Jingjing²

¹School of Information and communication Engineering
Beijing University of Posts and Telecommunications #186, 100876
Lee@bupt.edu.cn, zhangchuang@bupt.edu.cn, terrily@hotmail.com, wuming@sina.com

²International School
Beijing University of Posts and Telecommunications
Beiqijia Town, Changping District, Beijing, 102209
xiejingjing113@gmail.com

ABSTRACT
Crowdsourcing is applied more widely in many areas. However, the quality control method still needs future improvement. A new quality control method is proposed through worker’s experience in the work which has been divided into several stages. Workers in each stage are permitted to work on a number of HITs in proportion to their estimated accuracy in previous stages. To test the method, two experiments are conducted on CrowdFlower, and a simulation model is created based on Gaussian distribution and worker quantitative distribution in some existing crowdsourcing result data. The accuracy of result has increased from 76% to 85% in the first experiment, and in the simulation the accuracy of result has increased from 79.75% to 91.5% in simulation program.

Categories and Subject Descriptors
H.3.4 [Systems and Software]: Performance evaluation (efficiency and effectiveness)

General Terms

Keywords
Crowdsourcing, search evaluation, quality control, worker experience.

1. INTRODUCTION
In recent years, the emergence of crowdsourcing provides a new solution for many areas. The use of crowdsourcing platforms such as Amazon Mechanical Turk (http://www.mturk.com/) and CrowdFlower (http://crowdflower.com/) can provide a large number of online labors to complete the task quickly and inexpensively.

However, crowdsourcing is facing a new problem. Compared to traditional workers, we lose control of the workers online. They are inconvenient to communicate with and supervise so that the quality control of their work becomes difficult to be implemented. Therefore, there are always a considerable number of sloppy workers in crowdsourcing platforms performing random click and adversarial behavior, and it’s very difficult to detect them during crowdsourcing work.

Currently, there are two methods of quality control in crowdsourcing: control by consensus algorithm, and control by monitoring workers’ behavior. The effect of consensus algorithm is limited if the overall quality of crowdsourcing work is not good enough, and workers’ behavior is not an accurate representation of their ability. So a method to control quality through the history work record of workers is proposed.

Our strategy is: firstly, divide a crowdsourcing work into several stages in the first; secondly, find and record the sloppy workers and professional workers in the first few stages, and then prevent sloppy workers from working or restrict them in the next stages and encourage professional workers to do more work. The method can be used in various kinds of crowdsourcing works such as search and relevance evaluation.

The paper is organized as follows: in section 2, some related works and the current methodology to control quality of crowdsourcing are reviewed. Our strategy is described in section 3 and the experiment produces are in section 4. In section 5 the simulation model and the data which simulation is based on are shown. Our results analysis and directions for future work are summarized in section 6 and section 7.

2. RELATED WORK
Before the concept of crowdsourcing, an expectation maximization algorithm using maximum likelihood was presented, for inferring the error rates of annotators that assign class labels to objects, when the “gold” truth is unknown. The EM algorithm of them takes a set of objects as input, each being associated with a true class label and annotated by some workers. The EM algorithm iterates between estimating the correct labels for each of the objects, and estimating the error rates for each worker.[2] Based on the algorithm, some other algorithms have been presented, to reduce the impact of the sloppy workers in task and obtain high quality judgments.[4]

There is an algorithm to estimate the quality of the workers more accurately, using a confusion matrix. It estimates the cost matrix to consider the misclassification costs when an object of class A is classified into category B. The algorithm decreased the cost of annotation by 30% in a crowdsourcing experiment, while increasing the quality of annotation from 0.95 to 0.998.[5]
Another experiment includes five tasks using Amazon’s Mechanical Turk system. A technique for bias correction is proposed that significantly improves annotation quality on two of five tasks. It is proved that many large labeling tasks can be effectively designed and carried out in this method at a fraction of the usual expense.[6]

Besides consensus algorithm, the other option of quality control is to control by the workers’ behavior, to detect, correct and filter the sloppy workers. In a local search relevance judgments collection through crowd sourcing using Amazon's Mechanical Turk system, only 30% of the labels from the best workers was kept. Keeping the data originating from 30% of the best judgments leads to the same ITA (inter-annotator agreement) as what was obtained from trained labelers.[7]

There is a two-step approach being practiced. The first step was the pilot that consisting of a single HIT involving one video would be used for the purposes of recruiting and screening workers, and the second step was the main task. The pilot contained some questions about workers’ background, habits and more important, work attitude. The workers are chosen for the main task from the participants of the pilot by considering the quality of their description and choosing a diverse group of respondents. Crowdsourcing works done by these workers are reliable with high quality.[8]

A study analyzes the behavior of assessors that participated to identify some patterns that may be broadly indicative of unreliable assessments. Time analysis and the trap questions (the answer is known and very easy) are used to find out the sloppy workers.[9]

A good algorithm can refine the judgments and achieve consensus in crowdsourcing, but it still needs the professional workers to provide high quality judgments. In other words, the effect of consensus algorithm is limited by the overall quality of crowdsourcing work. The other option, worker control, is usually based on the behavior of workers, such as the time spent. But the behavior is not a very accuracy representation of worker capacity. It’s why our method is proposed in next section.

3. ASSIGNMENT DISTRIBUTION THROUGH WORKERS EXPERIENCE
Our strategy is to control quality by the worker control, but not based on workers’ behavior. Obviously, the accuracy of worker is a more direct representation than behavior. Therefore, we divide the whole task of crowdsourcing in to several stages so that workers will be restricted in stages based on the accuracy of them in the previous stages.

We use a level system to control workers through their experience, in which workers are divided into 3-5 levels based on their accuracy in previous stages. The HIT number the workers can do in a stage is limited to different number in different level. In the high level, the workers are encouraged to take more HIT so the limit number is high and in the low level just the opposite. The level system is more flexible than the dichotomy in which workers are divided into just two levels; can take HIT or not.

In order to obtain the accurate rate of workers accurately, the EM algorithm [9] is used in each stage. Specific procedures are as follow:

1. Divide crowdsourcing task into N stages, and clear the worker record.
2. Start the stage S_n, publish the HITs of the stage on crowdsourcing platform.
3. For each worker W_i, search the worker ID in worker record.
   3.1. If it is in the record, limit the HIT numbers the worker can do in stage S_n to the limit number M_i.
   3.2. If it is not in the record, limit the HIT numbers the worker can do in stage S_n to the initial limit number M_{init}.
4. Process the result data back from crowdsourcing platform with the EM algorithm [3], and store all new worker IDs in the worker record.
5. For each worker W_i, compute the correct rate of workers G_i based on the algorithm result, and update the HIT limit number M_i in worker record based on G_i.
6. Repeat the 2-5 steps until the whole work is finished.

In this method, the effect of sloppy workers is confined in their first stage because their HIT numbers are limited in the follow-up stages. If set the HIT limit number M_i = 0 when the correct rate is lower than a certain threshold, the sloppy workers will be refused to get more HITs.

Normal workers and professional workers are restricted in their first stage, but their limits are relaxed in the follow-up stages through their own effort. Similarly, if set the HIT limit number M = \infty when the correct rate is higher than a certain threshold, the limit of professional workers will be lifted.

The method needs more old workers (worker who has work record) than new workers (worker who has no record) because the old workers are limited based on their performance and the new workers are limited to the same level. Increasing the rewards of the old workers is a feasible way to encourage workers to participate in the following stages.

4. EXPERIMENT DESIGN
The experimental dataset is obtained from the 20 Newsgroups data. We selected 10 topics from the 20 Newsgroups. For each topic there are 20 documents. The annotation task on CrowdFlower is about to judge the relevance between the topic and document. Workers can mark the relevance with “Yes”, “No”, or “Unknown” (i.e. not sure).

As we said before, the method we designed requires the workers to participate in the stages as much as possible so that we can get the workers’ records to determine the extent of his HIT limitations in the work. We intended to encourage workers’ participation in more stages by increasing the rewards of the old workers, but our crowdsourcing platform CrowdFlower does not support this approach. So we decided to adopt another way to simulate the method in experiment. Specific procedures are shown below:

1. Collect K judgments per document without quality control on the CrowdFlower, and clear the worker record.
2. For each document D_i, select the first J completed judgments as the control group data of crowdsourcing without quality control.
3. Divide the whole work to N stages.
4. Start the stage S_n. Select the first J completed judgments per document.
5. For each worker W_i, search the worker ID in worker record.

1 The 20 Newsgroups data is available at: http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html
5.1. If it is in the record, limit the HIT numbers the worker can do in stage $S_n$ to the limit number $M_i$.

5.2. If it is not in the record, limit the HIT numbers the worker can do in stage $S_n$ to the initial limit number $M_{ini}$.

6. Reduce the judgments of workers exceed their HIT limit number and add the judgments of other workers in the chronological order of the HIT completion to make up the $J$ judgments per document.

7. Process the data selected through the EM algorithm and record all stage $S_n$ new worker IDs in the worker record.

8. For each worker $W_i$, compute the correct rate of workers $G_i$ based on the algorithm result, and update the HIT limit number $M_i$ in worker record based on $G_i$.

9. Repeat the 4-8 steps until the whole work is finished.

In our experiment, we collect 15 judgments per document without quality control on the CrowdFlower and select 5 judgments per document in 4 stages. i.e., we set $K = 15$, $N = 4$, $J = 5$. We set $K$ much more than $J$ in case that some judgments from former stage are reduced in Step 6. Our interface design is shown in Figure 1.

**Figure 1. The HIT interface on CrowdFlower**

5. SIMULATION DESIGN

The simulation model of crowdsourcing is designed based on the analysis of some existing data. By using the worker quantitative distribution of data, we determine the parameters of Gaussian distribution in the simulation model.

5.1 Existing Data Analysis

The data in the paper of Snow, O’Connor, Jurafsky and Ng. has 800 objects and each object has 10 judgments collected from AMT (Amazon Mechanical Turk). A HIT of crowdsourcing is 20 objects. They collect a total of 8000 judgments from 164 workers. Because the data includes ground truth, the accuracy rate is calculated easily.

In order to visualize the performance of workers, Figure 2 shows the accuracy rates on the vertical axis and the number of annotations on the horizontal for individual workers.

**Figure 2. The accuracy and the label number of workers**

The points in Figure 2 were clearly divided into three types: the left part is the normal workers part, which tend to take few HITs but work seriously; The lower right corner is the sloppy workers part, which tend to do a lot of HITs but low quality; the above of the central part is the professional workers part, which tend to provide results with higher quality than other workers.

5.2 Simulation Model

Based on the data, we design a simulation model which simulates the whole process of crowdsourcing.

Each worker $W_i$ in simulation has two parameters to control his activity: $G_i$ and $P_i$.

$G_i$ is the probability of that one judgment of worker $W_i$ is correct. When a document $D_j$ of true class label $T_j$ is labeled $L_{ij}$ by worker $W_i$, the $G_i$ is

$$G_i = P(L_{ij} = T_j)$$

Obviously, for the binary labels, the number of correct judgments in a HIT of worker is consistent with binomial distribution. For a HIT of $N$ documents, the number of correct judgments of worker $W_i$ is $N_c$. The probability distribution function of $N_c$ is

$$P(N_c = n) = C_n^k G_i^n (1-G_i)^{N-n}$$

(2)

$P_i$ is the probability of that worker quits crowdsourcing after finishing a HIT. When the number of HIT done by worker $W_i$ is $N_i$, the $P_i$ is

$$P_i = P(N_i = n | N_i \geq n)$$

(3)

The probability distribution function of $N_i$ is

$$P(N_i = n) = P(1-P_i)^{n-1}$$

(4)

The model includes three types of worker above: normal worker, sloppy worker and professional worker. Obviously, professional worker has a high value of $G$, normal worker has an average value which is consistent with Gaussian distribution, and sloppy worker has a value which represents random clicking. The normal workers have a lower probability $P$ to continue working and the sloppy workers and the professional workers have a higher one. Most workers are normal, and the amount of sloppy worker is a little
more than the professional. The parameter values and number of three types of worker are shown in Table 1.

Table 1. The parameters of three worker types

<table>
<thead>
<tr>
<th>Type</th>
<th>G</th>
<th>P</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>N(0.8, 0.000625)</td>
<td>0.9</td>
<td>88%</td>
</tr>
<tr>
<td>Sloppy</td>
<td>0.5</td>
<td>0.01</td>
<td>10%</td>
</tr>
<tr>
<td>Professional</td>
<td>0.95</td>
<td>0.05</td>
<td>2%</td>
</tr>
</tbody>
</table>

The results of workers in the model will be recorded when they complete the HITs. Records include the type of the worker, the number of HITs completed by the worker, accuracy and the HIT content. There is a probability of old workers (the workers with historical record) continuing the work of following stages. In our simulation, the probability is 0.5.

To simulate the experiment above, specific procedures of the simulation program are as following and the program flow chart is shown in Figure 3:

1. Divide crowdsourcing task to N stages, and clear the worker record.
2. Start the stage $S_n$, collect the worker judgments of the stage in the simulation program.
3. Each worker should not start their work until they pass the qualification test of $N_q$ documents which requires $N_{qc}$ documents are correctly labeled.
4. Process the result data back from crowdsourcing platform with the EM algorithm and record all new worker IDs in the worker record.
5. For each worker $W_i$, compute the correct rate of workers $G_i$ based on the algorithm result, and update the HIT limit number $M_i$ in worker record based on $G_i$.
6. Repeat the 2-5 steps until the whole work is finished.

In our simulation, we collect 5 judgments per document in 4 stages and the qualification test requires 3 correct labels of 5.

6. RESULT AND ANALYSIS

We conducted two experiments on crowdsourcing platform CrowdFlower. In the first experiment, no quality control of CrowdFlower is used in the Step 2 of the procedures in Section 5, and in the second experiment we collected 20 judgments per document (the same data as the first one) with quality control method which is provided by CrowdFlower in Step 2. Both results are as follows as a comparison. We also simulate four different situations in simulation program, and the parameters and results are shown in the end.

6.1 First Experiment Result

We use 3 different parameters of how to set the value of HIT limit number based on correct rate, to compare the advantages and disadvantages between them. For comparison, the control group data (without HIT limit) also is divided to 4 stages. The result is shown in Table 2 and Figure 4-11.

It can be seen from the results that, the accuracy of majority vote and EM algorithm in assignment distribution through workers experience (ADWE for short in following paper) are both lower than control group in the Stage 1. It is because ADWE limits the three types of workers to a same number. When the workers are in worker record, they are limited to the different number based on their correct rate in the Stage 2-4. So the accuracy of majority vote, EM algorithm and ROC value have all increased in the experiments, compared to the control group without quality control. The highest values of accuracy which can be obtained using parameter 2 or 3 are the same—85%.
Figure 4 shows the accuracy and the label number of workers in the experiment. We can obtain that the average accuracy of them is 63.839%, which is little more than control group because there is no worker control in control group. The average accuracy of workers rises from 60.1% in control group to 72% in experiment group with parameter 2. The results of three parameters are compared in Figure 11.

Figure 4. The accuracy and the label number of workers in experiment

Figure 5. The accuracy comparison between experiment result with parameter 1 and control group

Figure 6. ROC Curve of Parameter 1

Figure 7. The accuracy comparison between experiment result with parameter 2 and control group

Figure 8. ROC Curve of Parameter 2

Figure 9. The accuracy comparison between experiment result with parameter 3 and control group

Figure 10. ROC Curve of Parameter 3
6.2 Second Experiment Result

The second experiment is on CrowdFlower with the quality control provided by the platform. The result is shown in Table 2 and Figure 12-13.

Table 2. The result of experiment with quality control

<table>
<thead>
<tr>
<th></th>
<th>Majority Vote</th>
<th>EM Algorithm</th>
<th>Average Accuracy</th>
<th>ROC of MV</th>
<th>ROC of EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Group</td>
<td>99%</td>
<td>99%</td>
<td>92.5%</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>Parameter 2</td>
<td>99%</td>
<td>99%</td>
<td>94.4%</td>
<td>≈1</td>
<td>≈1</td>
</tr>
</tbody>
</table>

Through the quality control mechanism of CrowdFlower, the probability of low quality workers significantly reduces in the Figure 12 and the average accuracy of workers increases to 93.762%. The EM algorithm accuracy of the experiment and the control group are both 99.043% because CrowdFlower has provided a good-enough quality control. Therefore, we only provide the result of experiment with parameter 2. However, the average accuracy has a large increase. The 94.45% of experiment with parameter 2 is higher than the 92.476% of control group with quality control, and it can reach 96.27% with parameter 3, as is also shown in Figure 13.
6.3 Simulation Result
We simulate four different situations in simulation program, and they are shown in Table 2. We collect 4000 judgments of 800 documents in each situation of simulation. The result of simulation is shown in Table 2 and Figure 14-15.

The accuracy of experiment divided into stages is much higher than the control group in the figures and the accuracy of experiment divided into stages with qualification test can reach up to 91.5%.

Since the control group was not divided into stages in simulation, the accuracy of control group reflects the average value of whole work so its value is stable in all stages.

![Majority vote accuracy comparison between four simulations](image1)

**Figure 14.** The majority vote accuracy comparison between four simulations

![EM algorithm accuracy comparison between four simulations](image2)

**Figure 15.** The EM algorithm accuracy comparison between four simulations

7. CONCLUSIONS AND FUTURE WORK
In this paper we have proposed a method through the history of the workers to control the quality of crowdsourcing work. Two different experiments are conducted on the CrowdFlower. We analyzed some existing data and created a simulation model on the analysis result. Using the simulation model and real experiment, we tested this method and proved the effectiveness. The accuracy increased from 76% to 85% in first experiment, and the average accuracy of workers was increased from 92.5% to 96.27%. In simulation, the accuracy can reach up to 91.5% and the accuracy of control group is 79.75%.

Our method has several advantages. Firstly the sloppy workers are detected, recorded and restricted in our crowdsourcing work. Determining the sloppy workers by the work history of them is more accurate than by the behavior of them. Secondly, the professional workers are encouraged to do more HITs in crowdsourcing to improve the quality of work. At last, compared to filter the data after the completion of crowdsourcing work, our method prevents the sloppy workers from finishing a lot of labels which are filtered in the former method, in this way to save the money and time.

Our further research will focus on a method of quality control by considering the worker history and the worker behavior together. Another focus is a new consensus algorithm which takes the difficulty of the object labeled into account. Based on them, we will propose a more efficient method of quality control.

8. ACKNOWLEDGMENTS
This research is supported by “the Fundamental Research Funds for the Central Universities”. Program of China under GRAND NO. BUPT 2009RC0128, the 111 Project of China under Grant NO.B08004.

The research leading to these results has received funding from CrowdFlower. We would like to thank CMU for the 20-Newsgroup data that we used to run our experiment. We also thank Matthew Lease and Catherine Grady, Carsten Eickhoff, and Rion Snow, for providing us with their data and result.

9. REFERENCES


