Application research of iterative detection strategy in the crowdsourcing quality evaluation

Zhiyun Zheng, Xiaoqiang Guo, Zhenfei Wang, Dun Li
School of Information Engineering, Zhengzhou University, Zhengzhou 450001,China
E-mail:iezfwang@zzu.edu.cn

Quanmin Li
Tourism information center of Henan province, Zhengzhou 450001,China

***Abstract*—Crowdsourcing appeared on the Internet as a new service mode. In order to improve the accuracy of crowdsourcing results evaluation, it presents a crowdsourcing iterative detection strategy. According to the task crowdsourcing workers complete, we can use the principle that the minority is subordinate to the majority to assess the results of the task. The result sets of the assessment task in which option is not unique will be regarded as a new task to release in the crowdsourcing platform. Candidate workers will be chosen to participate in the iterative detection operations until the optimal result of each task has been determined. Experimental results show that compared with the classic quality assessment algorithm based on entropy, this mechanism can achieve better results.**

Keywords—crowdsourcing; quality control; entropy; iterative detection strategy; quality evaluation

I. INTRODUCTION

Crowdsourcing is put forward by the wired magazine reporter Jeff Howe for the first time in 2006, 6 article[1]. The concept is used to describe the process of distributing of work on the Internet, which found that creative or solve the technical problems of new business model. It has been appeared crowdsourcing platform at domestic and abroad, such as CrowdFlower and Amazon MTurk. Enterprises and organizations can use the originality and ability of freelancers to solve the problem by the crowdsourcing platform. These free workers who have the skills to complete the task will work in the spare time and obtain a certain reward. Typical crowdsourcing model is shown in figure 1. With the continuous development of crowdsourcing technology, it has been widely used in many fields, for example, in the field of information retrieval image search [2], data mining [3], the credibility of the microblog information calculation [4], CrowdDB query system in the field of database research [5]. In the aspect of sensor, Demirbasetal [6] proposed a sensing system based on the crowdsourcing, all tasks can be distributed to the online network users in the form of crowdsourcing.

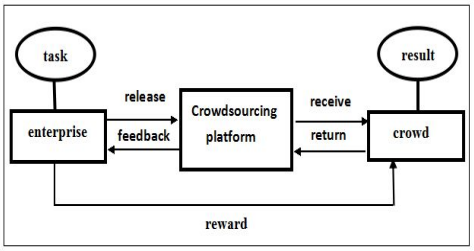

Fig. 1. A typical model of crowdsourcing.

The task of crowdsourcing release is geared to the needs of all users on the Internet. Because task workers have anonymous identity and different attitude, it will lead to greater uncertainty of the results of crowdsourcing task. In order to solve these problems, we put forward a crowdsourcing iterative detection strategy in this paper. We use the principle that the minority is subordinate to the majority to assess the results of the task, which can effectively identify the similar results existing in the assessment tasks. The tasks of similar results will be regarded as the new task released to crowdsourcing platform. Then they will select candidate workers to participate in the iterative detection. Thus it improve the accuracy of crowdsourcing quality evaluation. The paper is organized as follows: it introduces the quality control of related research work in the section II, it propose the architecture of crowdsourcing iterative detection strategy in section III, experiment and result analysis will be argued in section IV, summary and outlook in section V.

II. RELATED RESEARCH

With the wide application of crowdsourcing technology, the quality control of crowdsourcing get more and more attention of people, just as Lease Matthew pointed out that [7], if you care about the quality of the data, we must consider the problem of quality control. Now the quality control of the research work mainly focuses on three aspects[8]: (1) results quality evaluation method research. evaluate results submitted by the workers to identify malicious workers through a variety of methods; (2) the organization model of workers. control the quality of the crowdsourcing results by the establishment of a good worker organization management mode; (3) the design of crowdsourcing task. obtain high quality results from the perspective of designing a good crowdsourcing task goal. This

article mainly aims at the results of crowdsourcing quality evaluation method, the method is described below.

A. Gold standard data evaluation strategy

Gold standard data is one of the common methods for quality evaluation currently. It can detect the cheating type workers by comparing the submit results with the standard answer and refuse their answers. For some simple tasks, it is an ideal choice for gold standard data, however, there is no fixed answer for the most tasks released on the crowdsourcing platform, a lot of problems should be evaluated objectively and fairly. Therefore, it is very difficult to use the gold standard data to evaluate methods for this kind of problem.

B. Dynamic hierarchical crowdsourcing quality control strategy

Different from gold standard data evaluation strategy, reference [8] proposed staged dynamic crowdsourcing quality control strategy. The method implemented stage type dynamic quality control. it designed several segmented check points in the process of the completion of crowdsourcing tasks, if finds the result of low quality on a stage, it will stop the workers to take part in the mission and delete the submitted result, then new workers will be chosen to continue this phase of the mission. Using this strategy can be find the cheating type workers sooner and reduce the final result in unreliable results proportion, finally improve the overall quality of the results.

Although the staged dynamic crowdsourcing quality control strategy can improve the results quality of crowdsourcing task, but due to set up the testing point, it need take more time for the results inspection and replacement after completion of each stage task, which will extend the task completion time; at the same time, how to reasonably set up check points and replace the strategy must be seriously considered, if the replacement strategy is designed unreasonably, it may fall in a vicious cycle, and bring serious impact on the completion of tasks of crowdsourcing.

C. The crowdsourcing quality evaluation algorithm based on entropy

Expectation Maximization Algorithm[9] (maximum expected estimation) is a classical crowdsourcing quality assessment Algorithm. Ipeirotis P G et al. [10] by this Algorithm adopt the form of a matrix output. Due to the situation that is not intuitive show workers to complete the task, the literature [11] proposes a crowdsourcing quality evaluation algorithm based on entropy. The algorithm introduces definition of entropy , all the workers will be calculated results after completing the task, which can draw every worker's score, the good worker's score is close to one, and the poor worker's score is almost zero. By worker's score, it can directly show completing the task. Evaluation algorithm based on entropy calculate the unknown parameters by EM algorithm, but when the EM algorithm adopts the principle that the minority is subordinate to the majority, selects the most options as the best results and ignores similar results, it will lead to a decrease of accuracy of the evaluation result.

III. THE CROWDSOURCING QUALITY EVALUATION ARCHITECTURE BASED ON ITERATION

According to the insufficient quality control work in the current crowdsourcing research, we put forward a kind of architecture based on the crowdsourcing quality evaluation. By analysis the result which workers submit, we use iterative control strategy (iterative control strategy) to identify tasks existing the similar result set and compute the accuracy of the results which workers submit, then keep good crowdsourcing task results.

A. The crowdsourcing quality evaluation architecture

Figure 2 depicts the crowdsourcing quality evaluation architecture, mainly includes three parts: crowdsourcing task distribution, workers classification and iterative detection strategy. Crowdsourcing task distribution is responsible for the task in the pool released to the crowdsourcing platform; Worker classification algorithms will chose excellent workers to join the candidate crowd; Iterative detection strategies to evaluate the similar results exists the result set by iterative operation to establish the best results.

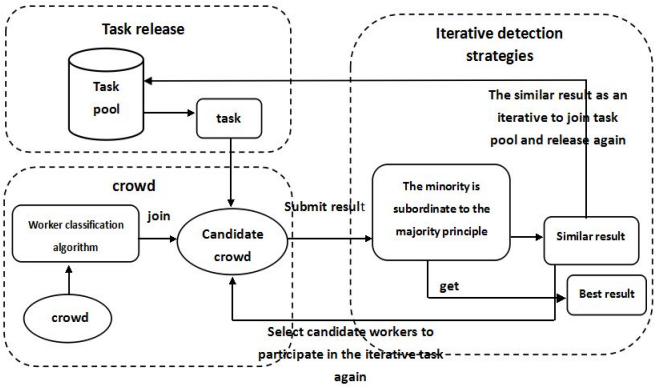

Fig. 2. Crowdsourcing quality evaluation architecture.

B. Worker classification algorithm

Workers who participate in the task of crowdsourcing are usually divided into the following categories: (1) diligence types of workers. These workers who listen to instructions and are strictly in accordance with the requirements of the task are often given the satisfactory results; (2) the hasty workers. May have good intentions, but there is no serious reading topic leading the results of the low quality; (3) the workers of the malicious. They often submit the results of question by deception.

Iterative detection strategies screen the crowdsourcing task with similar results and release again in the process of each iteration, then select candidate workers to complete. Reference [12] gives workers classification method, which can detect the malicious and hasty workers.

For detecting premature types of workers, we adopt the method of correlation characteristic distance, which can calculate random points of each worker by comparing the results with other workers.Using the following type (1) to calculate random points.

$$RandomSpam = \frac{\sum_{j \in J_w} \sum_{i \in J_{j,\overline{w}}} dis_{ij}}{\sum_{j \in J_w} | J_{j,\overline{w}} |} \quad (1)$$

w stands for workers, j_W stands for relevance judgment belongs to the workers w do, $J_{j,\overline{w}}$ stands for in addition to the workers w collection correlation judgment, other workers do. dis_{ij} said that for the same problem j, workers w and other workers i do judgment between difference distance, if dis_{ij}=0, both do the same. On the other hand, if dis_{ij}=1, both do different judgment.

For checking malicious types of workers, we will use the number of inconsistent of complete results with other workers. calculate the score of each worker through the type (2) .

$$UniformSpam_W = \frac{\sum_{s \in S} |s| \cdot (f_{S,J_w} - 1)\{ \sum_{j \in J_{s,w}} \sum_{i \in J_{j,\overline{w}}} (disagree_{ij})^2 \}}{\sum_{j \in J_w} | J_{j,\overline{w}} |} \quad (2)$$

S is a collection of all tasks. $disagree_{ij}$ said that for the same problem j, workers w do relevance judgment $J_{s,w}$ not agree with other workers to submit. $f_{s,Jw}$ said that the frequency of workers w tag tasks in its judgment J_w.

According to the experimental verification, score selected by formula 1 is greater than 0.7, score selected by formula 2 is greater than 1.6, it can effectively find the malicious types of workers existing in the crowd .

C. The minority is subordinate to the majority principle

The task evaluation standard is based on the principle that the minority is subordinate to the majority among the crowdsourcing quality evaluation architecture. The principles are defined as follows: supposing that one crowdsourcing task, m workers participation, task of candidate sets H have s option ($h_1 \sim h_s$), benchmark set for $k = \lfloor m/s \rfloor$, threshold value between similar options is set for δ , the choice of each option number respectively $c_1, c_2, \ldots, c_t$, ($c_1 + c_2 + \ldots + c_t = m$). If any option h_i and h_j, meet $c_i > k$, $c_j > k$ ($1 = <i, j = < t$), and $| c_i - c_j | < \delta$, argues that hi and hj is the result of approximate options; Otherwise, if $| c_i - c_j | > \delta$ and $c_i > c_j$, argues that option h_i option is optimal results. Now there are six workers participating in crowdsourcing tasks, the task candidate has two options, benchmark for 3, worker's choice of task results as shown in figure 3:

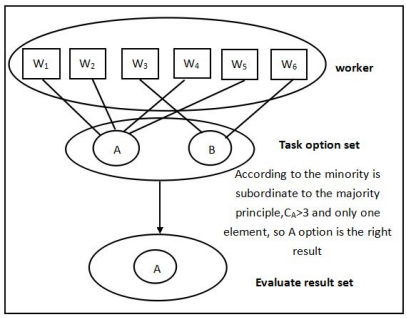

Fig. 3. The minority is subordinate to the majority principle model diagram.

D. Iterative detection strategies

It proposed an iterative detection strategy in the paper, its essence is preprocessing operations for the results of crowdsourcing task, it will filter out the results of low accuracy. The basic idea of the iteration strategy: Firstly, according to the result of initial crowdsourcing collection R_{m*n}, section III presents the principle that the minority is subordinate to the majority to eliminate the most impossible right results; Then the possible similar results in task regard as the new task released on crowdsourcing platform and select candidate workers participate in the task, which adopt the principle of the minority is subordinate to the majority again to screen the result set and approximate calculation. Multiple iterative operation can gradually achieve the accuracy of the result and ultimately determine the best result of each task; Finally compare results set $G_n = \{g_1, g_2, \ldots\ldots, g_n\}$ with the initial result set $R_{m \times n} = \begin{vmatrix} r_{11} & r_{12} & \ldots\ldots & r_{1n} \\ r_{21} & r_{22} & \ldots\ldots & r_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ r_{m1} & r_{m2} & \ldots\ldots & r_{mn} \end{vmatrix}$,we can calculate the accuracy rate of each worker (w_i) , as follows:

$$rate\ [i] = \frac{\sum_{j=1}^{n} R_{ij}}{\sum_{j=1}^{n} G_j} \quad (3)$$

Iterative detection algorithm is given below:

Input: crowdsourcing task T= {$t_1, t_2, \ldots, t_n$}, task option set H_i={$h_1, h_2, \ldots, h_s$} (for i=1,....n), evaluate result G ={null}

Output: the evaluate results G ={$g_1, g_2, \ldots\ldots, g_n$}, the accuracy rate of workers (w_i)

Initialization: according to worker classification algorithm, select workers={$w_1, w_2, \ldots\ldots, w_m$}, benchmark $k = \lfloor m/s \rfloor$, threshold value δ , initial result set $R_{m \times n}$ =null

While ($|G| <$ n)

{ S1: issue task T and H_i set to crowdsourcing platform, m workers participate in the task;

S2: according to the minority is subordinate to the majority principle to calculate result sets R, adding task t_i the best result to G;

S3: update task T, the H_i of task t_i, and evaluation result G;

}

End

IV. EXPERIMENTS AND RESULT ANALYSIS

In order to demonstrate the effectiveness of crowdsourcing iterative detection strategy, we build a crowdsourcing experiment platform. Experimental scheme is as follows: it release a set of tasks combining with the specific teaching environment on the crowdsourcing platform and select 20 candidate students to participate in the task. We use two different methods of quality evaluation for the same set of tasks in the process of experiment. The first methods integrate the iterative detection strategies into crowdsourcing experiment platform and perform the crowdsourcing mission. Because of existing similar results in the experiment, we should set up a reasonable threshold and calculate the evaluation results; Then we use the classic quality assessment algorithm based on entropy to evaluate crowdsourcing results; Finally we use quality assessment results based on the entropy as a benchmark, compare it with the results of the two kinds of algorithm .

A. Experimental environment

Hardware: the dawning 1420r-G server, 32GB of memory, faster 3.07 GHz; Software: MyEclipse9.0

B. Analysis of experimental results

We release a set of crowdsourcing tasks that include 50 tasks on crowdsourcing platform. The content of the tasks involved are computer related professional knowledge. Each task have five options, with the 'A', 'B', 'C', 'D' and 'E'. We choose 20 students from candidate crowd to complete the tasks of crowdsourcing. Table 1 shows that the 20 students participate in crowdsourcing task for the first time after the completion of the partial results of statistics.

TABLE I. WORKER OF CROWDSOURCING PARTIAL RESULTS OF STATISTIC TASK

option	A	B	C	D	E
Task one	3	7	2	2	6
Task two	3	3	2	7	5
Task three	7	3	2	4	4
Task four	2	2	5	5	6
Task five	5	3	9	1	2

a) Setting threshold.

In the experiment, how to set up a reasonable threshold is the most critical, it is decided by the number of people involved in crowdsourcing. For example, data in table 1, we use the principle that the minority is subordinate to the majority, set the benchmark for 5, namely as long as the option of the task has more than five workers, this option may be the right result; Threshold is set to 2, that is, if the result of the absolute value of the difference between in the range of 2, these options are similar results. We can see B and E are the correct results for task 1 , considering the threshold value is 2, the absolute value of B and E are in the setting of threshold value range, so B and E are similar results B in the task 1. To task 1, B and E results will be regarded as the new crowdsourcing task released on crowdsourcing platform and select the candidate of workers to participate in the iteration. Similarly, to task 2, D and E are probably right results, the absolute value of the difference between the two options within the scope of the threshold, so D and E are the close result ; Task 3 can determine A is correct result; To task 4 , C, D and E could be the right result, the absolute value of the three options are within the scope of the threshold value, the difference between the C, D and E are similar results; Task 5, A and C could be correct results, due to the difference between the two options of the absolute value is beyond the scope of threshold value, therefore, option C is the most correct answer.

For crowdsourcing results existing in the uncertain task, it just need the possible result set of task released on crowdsourcing platform, which choose the workers from the workers participate in the task of crowdsourcing iteration is complete. Below 20 students to participate in the task, for example, the threshold is set to 2, 3 and 4 respectively, iterative detection strategy is used to evaluate the results. The experimental results are shown in figure 4:

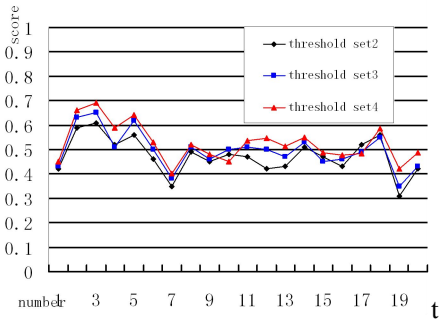

Fig. 4. The threshold setting contrast.

It can be concluded from the experimental results, the rationality of the setting threshold results will directly affect the accuracy of the assessment, if the threshold is too large there may be many similar results, although it can enhance the accuracy in crowdsourcing tasks, but it will increase the number of iterative detection too; On the other hand, if the threshold is too small and similar results will be screened out which lead the low accuracy of crowdsourcing task. For example, the task 5 in table 1, A and C are may correct results. When the threshold is set to 2, and 3, there is no similar results, after an iterative detection C will be determined to be the correct answer;When the threshold is set to 4, A and C is the result of the close, the need for the next iteration operation to be sure, but in fact the correct results in the second iteration is still C, due to the large threshold settings, make two similar options as far as A result, lead to the number of iterations more, in fact, the iterative operation is futile, extra overhead costs.

After experimental verification, this experiment is a reasonable threshold is set to 3.

b) The comparison of algorithm

Respectively use the iterative detection strategy and quality assessment algorithm based on entropy to evaluate the results of the 20 students complete, threshold to 3, the experimental results are shown in figure 5.

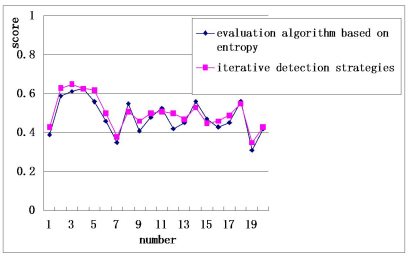

Fig. 5. Results comparison algorithm.

By analyzing the results of the experiment, it can be concluded that compared with the quality assessment algorithm based on entropy, iterative detection strategies fully considering that when adopting the principle that the minority is subordinate to the majority to evaluating quality, there may be a similar result in the same task. It needs iterative process to determine the best results, so as to improve the accuracy of evaluation results. However, when quality assessment algorithm based on entropy calculates the score of every worker, for involving the unknown parameters calculated using the EM algorithm, the algorithm initialization parameters adopt the principle that the minority is subordinate to the majority, just select one of the most options as the best quality assessment result. For example, for the task 1 in table 1, the best answer is E, but due to the malicious participate in make option B the number of workers increased. So the EM algorithm take B as the best result, and E is a relatively similar result and easily ignored, at this point, if we think the correct answer is B it will affect worker's score and decrease the accuracy of the evaluation results. Iterative detection strategies give full consideration to the above situation and deal with similar results thus improves the accuracy of crowdsourcing evaluation.

V. SUMMARY AND OUTLOOK

Aiming at the shortcomings of the current crowdsourcing quality control, it proposes an iterative detection strategy. According to the crowdsourcing workers task, we use the principle that the minority is subordinate to the majority to assess the results of tasks.The result sets of the assessment task in which option is not unique will be regarded as new task to release in the crowdsourcing platform. Candidate workers will be chosen to participate in the iterative detection operations until the optimal result of each task has been determined.The key points of the iterative detection strategy is to set a reasonable threshold. By comparing iterative detection strategy with quality assessment algorithm based on entropy, experiments results show that the iterative detection strategy is better than the quality evaluation based on entropy, in accordance with the expected result.

Iterative detection strategy is proposed to improve the accuracy of crowdsourcing results in the paper. However, most of the crowdsourcing task need to make the remuneration to workers, therefore, the next step of this paper will consider economic factors of crowdsourcing, namely under the premise that how to ensure the quality of crowdsourcing task results to make crowdsourcing publishers pay the least amount of compensation.

ACKNOWLEDGMENT

This paper is supported by the International Science and Technology cooperation Fund of Henan province with Grant No. 144300510007.

REFERENCES

[1] Howe Jeff. The rise of crowdsourcing. Wired, 2006, 14(6):176-183

[2] Yan Tingxin, Kumar V, Ganesan D. CrowdSearch: Exploiting crowds for accurate real-time image search on mobile phones//Proceedings of the International Conference on Mobile Systems, Applications, and Services.San Francisco,USA,2010:77-90

[3] Lease M, Carvalho V R, Yilmaz E.Crowdsourcing for search and data mining. Journal of SIGIR Forum(SIGIR), 2011, 45(1):18-24

[4] Castillo C, Mendoza M, Poblete B. Information credibility on twitter//Proceedings of the WWW.Hyderabad,India, 2011:675-684

[5] Franklin M J, Kossmann D, Kraska T, et al. CrowdDB: answering queries with crowdsourcing[C]//Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, 2011: 61-72.

[6] M. Demirbas, M.A.Bayir, C.G.Akcora, Y.S.Yilmaz. Crowd-sourced sensing and collaboration using twitter, in: Proceedings of the IEEE International Symposium on “A World of Wireless Mobile and Multimedia Networks”,WoWMoM, Montreal, Canada,June 2010

[7] Lease M. On Quality Control and Machine Learning in Crowdsourcing[C]//Proceedings of Human Computation Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence.San Francisco,California,USA, 2011:97-102.

[8] Z.Q.Zhang, J.S Pang, X.Q.Xie, and Y.Zhou. Research on Crowdsourcing Quality Control Strategies and Evaluation Algorithm[J].Chinese Journal of Computers,2013,36(8):1636-1649.

[9] Dawid A P, Skene A M. Maximum likelihood estimation of observer error-rates using the EM algorithm. Applied Statistics,1979,28(1):20-28

[10] Ipeirotis P G , Provost F, Wang J. Quality management on Amazon mechanical turk [C]//Proceedings of the ACM SIGKDD workshop on human computation. ACM, 2010: 64-67.

[11] Raykar V C, Yu S. An entropic score to rank annotators for crowdsourced labeling tasks[C]//Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2011 Third National Conference on. IEEE, 2011: 29-32.

[12] Vuurens J B P, de Vries A P. Obtaining high-quality relevance judgments using crowdsourcing[J].Internet Computing, IEEE, 2012, 16(5): 20-27.