Survey on Crowdsourcing Worker Quality Evaluatuation

Nikhitha Prasad

Abstract—Crowdsourcing is a relatively new problem-solving and production model. In this distributed computing model, tasks are distributed by enterprises through the Internet and recruit more suitable workers to involve in the task to solve. Later, Jeff Howe coined the term "crowdsourcing" in 2006. Till then, a lot of work in crowdsourcing has focused on different aspects of crowdsourcing, like techniques for computation and performance analysis technical difficulties. In this survey the various techniques and approaches that are used for evaluating workers on crowdsourcing environment are covered.

Index Terms— Crowdsourcing, Crowdsourcing systems, Big data ,Quality control,Task fingerprinting,Replacement algorithm,Modelling annotators

1 Introduction

As a new distributed computing model, crowdsourcing allows to leverage the crowd's intelligence and wisdom toward solving problems. Crowdsourcing can be defined as Taking a task, once performed by employees and outsourcing it to an large network of people by a person or a company. Crowdsourcing is obtained by merging the terms "crowd" and "outsourcing".Some well known applications of the model include Amazon Mechanical Turk, Threadless, iStockphoto, user-generated advertising contests InnoCentive, the Goldcorp Challenge

The explosive growth and widespread accessibility of the Internet have led to surge of research activity in crowdsourcing. Crowdsourcing has now arisen as a novel way for tasks that can be easily performed by humans but remain rather tough for computers.

This paper presents a survey of various methods of evaluating the quality of workers in a crowdsourcing environment.Here we discuss five different method used for quality evaluation.

The first method describes *Evaluating Translation Quality* [1] demonstrate that the evaluation of translation quality manually is not as time consuming or as expensive as generally thought. An online labor market Amazon's Mechanical Turk pays people small quantity of bounty in the form of money to complete human intelligence tests – the tasks that are complex for computers but simple for humans

Second method prototyped a task fingerprinting system [2] that uses JavaScript and the jQuery library to monitor user activity on crowdsourcing market web pages. It demonstrated how workers' behavioral traces can be used to make inferences about worker's task performance, including identifying cheaters, estimating output quality, and predicting errors

The third method explored the use of Amazon Mechanical Turk (AMT) to determine whether reliable natural language annotations can be provided by non expert labelers. To understand tasks that would be sufficiently natural and learnable for non-experts five natural language were chosen, and which had gold standard labels from expert labelers. The tasks are: affect recognition, event temporal ordering, word similarity, textual entailment recoginition, and disambiguation word sense. Each task, AMT was used to annotate data as well as measure the quality of the annotations. This was done by performing comparison between them and with the gold standard (expert) labels on the same data.

In next method a simple crowdsourcing model for the study of replacement strategies and dynamic worker evaluation. A set of active workers that all execute similar tasks in sequence are focused by the model, and also on evaluations based on disagreement of workers. a family of worker replacement policies based on worker threshold parameters are discussed. An evaluation framework and metrics that capture how quickly the system obtains a pool of high accuracy workers. a rule of thumb for selecting the threshold to be used in the replacement policy.

The fifth method uses an online algorithm that estimates the annotator's expertise or the reliability, and decides how many labels to request per image based on who has performed the labeling of it. The model is sufficient enough to handle many types of annotations, and showed results on binary ie only two value, multiple valued, and annotations that are continuous-valued, and is collected from MTurk.

2 Crowd Worker Quality Evaluation Techniques

Here four methods are discussed that are used for worker quality evaluation. That Evaluates Translation Quality, Discriminative variation on Latent Dirichlet Allocation, Multi-Aspect Sentiment Model, Multi-grain LDA, and Weakly Supervised Joint Sentiment-Topic Detection.

2.1 Evaluating Translation Quality

Detailed submission guidelines can be found on the author resources Web pages. Author resource guidelines are specific

to each journal, so please be sure to refer to the correct journal when seeking information. All authors are responsible for understanding these guidelines before submitting their manuscript. An existing set of gold standard judgments of machine translation quality were taken from the Workshop on Statistical Machine Translation (WMT), which performs a yearly large-scale human evaluation of machine translation quality. Computational linguists who develop machine translation systems were the experts who produced the gold standard judgments. All judgments from the WMT08 German-English News translation task were recreated. The 11 different machine translation system's output that participated in this task was scored by ranking sentences translated relative to each other of their output.

Then non-expert Turker judges were evaluated by measuring their inter-annotator agreement with the WMT08 expert judges, and also by comparing the coefficient of correlation across the various rankings of the machine translation systems produced by the two sets of judges. Each item is redundantly judged by five non-experts Figure 1 shows the effect of combining experts' judgments on their agreement with non experts. By examining each pair of translated sentence, agreement is measured and then counting if two annotators both indicated that A > B, A < B, or A = B. Chance agreement was found out to be 1/3. The top line indicates the inter-annotator agreement that exists between WMT08 expert annotators, who in 58% of the time agreed with each other

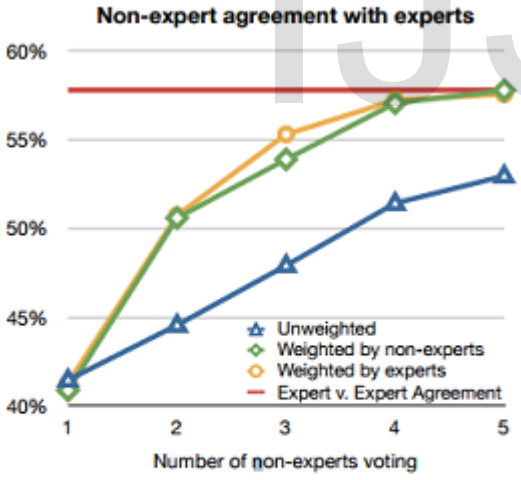

Figure 1: Agreement on ranking translated sentences increases as more non-experts vote. Weighting non-experts' votes based on agreement with either experts or other non-expert increases it up further. Five weighted non-experts reach the top line agreement between experts.

The low cost of the non-expert labor found on Mechanical Turk is cheap enough to collect redundant annotations, which can be utilized to ensure translation quality. By combining the judgments of many non-experts were able to achieve the equivalent quality of experts.

2.2Task Fingerprinting

Examining crowd worker's behavioral traces as they complete work can actually be used to predict the quality of their final product. Task fingerprinting technique is proposed to collect and analyze such behavioral traces in online task markets (this technique can also be generalized to other settings as well)

The quality of different workers was distinguished by Rzeszotarski and Kittur [2] by analyzing the behaviours of the workers. However, this method requires the crowdsourcing system to provide the worker's behaviour logs.

A task implies a worker performing some actions on an input (typically employer provides it) resulting in some output. The input can be a document to summarize, an image to tag or even just a set of guidelines for open response. The worker engages in a series of cognitive and motor actions that result in changes in their web browser using this input (e.g., mouse movements, scrolling, and keystrokes) and produces an end product for the requester.

Represented as:

$$F_{worker}(in_{task}) = out_{task,worker}$$

where, in_{task} : input
out_{task} : output

The input is given by the employer which is usually some sequence of motor and cognitive actions performed by the worker (f_{worker}) on the input, generating some output that is consumed by the employer.

In addition, information collected characterized the user's behavior in a holistic sense. Firstly, summary data was generated, like the total time the system was logging activity, the total amount of scrolling and mouse movement, the counts of different types of events, and the lengths of the event logs. These allows to get a general sense of what a user is doing in the environment.

Secondly, collect more specific information about the events, like the number of times a user pastes text, a total count of the number of unique keys a user presses, the number of time certain special keys like tab and backspace were used, and how many form fields were accessed. This information exposes users with especially unique behavioral patterns. Finally, collecting information about the delays the user introduces into their work. determine how long the user spent 'off focus' from the page, and the cumulative time they spent between keystrokes in a form field.We can use these features to make higher level judgments about user deliberation and attention in tasks

2.3 Training a system with non-expert Annotations

The Amazon Mechanical Turk system is employed in order to elicit annotations from non-expert labelers.The quality of non-expert annotations on five tasks: word similarity, affect recognition, temporal event recognition, recognizing textual entailment, and word sense disambiguation was analyzed. Then each annotation task and the parameters of the annotations using AMT were defined.

The evaluation of expert vs. non-expert labeler data annotations for five tasks found that for many tasks only a small number of non expert annotations per item are necessary to equal the performance of an expert annotator. In a detailed study of non-expert and expert agreement it required an average of 4 non-expert labels per item for an affect recognition task in order to emulate expert-level label quality.

2.4 Replacement Algorithms

Focuses on a set of replacement algorithms. It not only is the set intuitive and easy to describe, but it also allows a wide variety of alternatives by changing the values of a few parameters.

Replacement algorithm has two phases

(a) The state (or history) of each of the workers participating in the current round is updated

(b) The workers to be replaced are selected.

State: The application maintains the following state (or history) for each worker w:

- Number of participated rounds thus far, r(w)
- Number of rounds in which the majority opinion is agreed with w, c(w)

c(w)/r(w) is equivalent to the empirical accuracy of the worker, when the majority opinion is always correct. Another interpretation is that the fraction represents an estimate of the probability of the worker agreeing with the majority opinion.

Algorithm 1: Application Outline

```
Data: 𝒟, k
𝒜 ← ∅;
while |𝒜| < k do
    Add new worker from Worker Pool to 𝒜;
for t ∈ 𝒟 do
1   Ask workers in 𝒜 to evaluate task t;
2   Return v(t) inferred by task solving algorithm;
3   Evict workers in 𝒜 selected by replacement algorithm;
4   while |𝒜| < k do
5       Add new worker from Worker Pool to 𝒜;
```

Ineffective Worker Replacement: Algorithm uses the following rule to eliminate poorly-performing workers (p, rmin are parameters that we need to set): Replace w if c(w) r(w) < p ∧ r(w) ≥ rmin

Furthermore, while one simple space of algorithm has been explored, more general algorithms with more elaborate methods of judging the competency of workers (using weighted or time-decaying scores) can be considered, a more fine-grained record of history of workers, eviction of workers who have participated in a certain (large) number of rounds, eviction of a fixed number of workers in every round, and so on.

2.5 Modeling Annotators and Labels

An online algorithm was proposed to determine the ground truth value" of some property in an image from multiple noisy annotations. As a by-product it produces an estimate of annotator expertise and reliability. It selects which images to label based on the uncertainty of their estimated ground truth values, and the desired level of confidence.

The assumption that each image i has an unknown "target value" which is denoted by z_i . This may be a continuous or discrete scalar or vector. The set of all N images, indexed by image number, is I = {1…..N}, and the set of corresponding target values is abbreviated z = $\{z_i\}$ N i=1. The reliability or expertise of annotator j is described by a vector of parameters, a_j.

There are M annotators in total, A = {1……M}, and the set of their parameter vectors is a = $\{a_j\}$ M j=1. Each annotator j provides labels L j = $\{l_{ij}\}$ I ∈ I_j for all or a subset of the images, $I_j \subseteq I$. Likewise, each image i has labels $L_i = \{l_{ij}\}$ j ∈ A_i provided by a subset of the annotators $A_i \subseteq A$. The set of all labels is denoted L. For simplicity, assumed that the labels l_{ij} belong to the same set as the underlying target values z_i

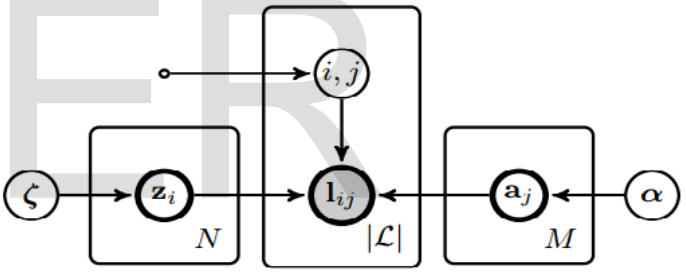

Figure 2: Plate representation of the general model. The i,j pair in the middle plate indicates which images each annotator labels, is determined by some process that depends on the algorithm

3 CONCLUSION

In this paper several methods used for analysis of worker quality evaluation in a crowdsourcing environment has been discussed. Crowdsourcing has a wide variety of applications in day to day life. The studies above are mostly focused on the traditional architecture and do not take the big data environment into consideration; thus the practicability and extensibility of these studies are not sufficient.

REFERENCES

[1] C.Callison-Burch, "Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk," in Proc. Conf. Empirical Methods Natural Language Processing, vol. 1, pp. 286–295,2009

[2] J. M. Rzeszotarski and A. Kittur, "Instrumenting the crowd: Using implicit behavioral measures to predict task performance," in Pro. 24th Annu. ACM Symp. User Interface Softw. Technol., pp. 13–

22,2011 Neural Information Processing Systems (NIPS), 2008 pp. 123-135, 1993. (Book style)

[3] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks," in Proc. Conf. Empirical Methods Natural Language ProcessingEmnlp, 2008, pp. 254–263.

[4] A. Ramesh, A. Parameswaran, H. Garcia-Molina, and N. Polyzotis, "Identifying reliable workers swiftly," Infolab Tech.Rep., Stanford Univ., Stanford, CA, USA, 2012.

[5] P. Welinder and P. Perona, "Online crowdsourcing: rating notators and obtaining cost-effective labels," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, pp. 25–32, 2010.

[6] D. C. Brabham, "Crowdsourcing as a model for problem solving:An introduction and cases," in Convergence the International Journal of Research Into New Media Technologies, SAGE, vol. 14, no. 1, pp. 75–90, 2008.

[7] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality control in crowdsourcing systems: Issues and directions," IEEE Internet Comput.,vol. 17, no. 2, pp. 76–81, Mar. 2013.

[8] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the world-wide web," . ACM, vol. 54, no. 4, pp. 86–96, 2011.

[9] P. Clough, M. Sanderson, J. Tang, T. Gollins, and A. Warner, "Examining the limits of crowdsourcing for relevance assessment," IEEE Internet Comput., vol. 17, no. 4, pp. 32–38, Jun. 2013.

[10] O.G. Staadt, B. Carpenter, "Multilevel bayesian models of categorical data annotation," [Online]. Available: http://lingpipe-blog.comlingpipe-white-papers, 2008.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," J. Roy. Statistical Soc. B, vol. 39, no. 1, pp. 1–38, 1977.

[12] P. Dawid, A. M. Skene, A. P. Dawidt, and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," Appl. Statistics, vol. 28, no. 1, pp. 20–28, 1979.