An Impact-driven Model for Quality Control in Skewed-domain Crowdsourcing Tasks

Kinda El Maarry
Institute for Information Systems
TU Braunschweig
Braunschweig, Germany
+49 (531) 391-3102
elmaarry@ifis.cs.tu-bs.de

Wolf-Tilo Balke
Institute for Information Systems
TU Braunschweig
Braunschweig, Germany
+49 (531) 391-7442
balke@ifis.cs.tu-bs.de

ABSTRACT

Not only do the highly-distributed digital crowdsourcing solutions surpass both borders and time-zones, but they materialize the vision of impact sourcing, by tapping into new labor markets in developing countries. Unfortunately, crowdsourcing is associated with severe quality issues. To that end, many countermeasures have been designed to detect spammers, except in practice, also honest, yet not perfect workers will often be exposed and deprived of much-needed earnings. Here, we argue for the need of an impact-driven quality control measure, especially for skewed-domain tasks. Such a measure should ensure high quality results, while simultaneously fulfilling the social responsibility aspect of crowdsourcing.

CCS Concepts

• Information systems ~ Collaborative and social computing systems and tools

Keywords

Crowdsourcing; impact sourcing; quality control; fraud detection

1. INTRODUCTION

With crowdsourcing emerging as an unprecedented international agile work force, companies are encouraged to hire this readily available workforce for intelligent information processing skills: content annotation [1], information extraction [2], sentiment analysis [3], etc. Unfortunately, the very nature of crowdsourcing's virtual workspace also encourages spammers to cheat the system for quick monetary gains; a random-answering mechanism may suffice and may be indeed very profitable. However, such cheating exposes the entire task to severe quality problems. Moreover, the countermeasures that are often deployed to fend off spammers, e.g. gold questions, reputation-based systems, majority voting, etc., tend to be discriminative and expose low-skilled honest workers as well. Actually, this is more severe than it sounds, since the social aspect of the crowd sourcing solution can be immense, with 1.8 billion people unable to access a formal job and half of the world's population living on less than $2.50 a day[1]. Thus, the need for an impact driven model is of importance.

Furthermore, designing such a quality model becomes even more complicated, when answer sets for crowdsourcing tasks are inherently *skewed,* with one answer being very prevalent [4]. This is indeed often the case with zipf-distributed web data. With such a setup, strategic spammers can easily exploit the inherent skewness to avoid detection, by always submitting the frequent answer. The simple, yet effective idea of such a strategy is to get highly accurate results. At which point, commonly used quality measures fail to identify strategic spammers, since they are outwardly doing a very good job, and in many cases, better than the honest workers. Next, we give a short overview of the related work. In the third section, we present some motivational case studies, which illustrate the problems of current countermeasures and how they discriminate against low-skilled honest workers. Finally, we conclude with the central design aspects that an impact-driven quality model should possess.

2. RELATED WORK

A lot of focus has been given to devising countermeasures against spammers. In this section, we briefly take a look on the most commonly used countermeasures. First, *gold questions*, are usually covertly added to tasks, to catch spammers off-guard. Upon failing to correctly answer a certain percentage of gold questions, the worker is declared to be fraudulent and is accordingly discarded from the workforce. Unfortunately, gold questions can only be utilized for factual tasks and can't be employed in opinion-based tasks or individual perceptions and sentiments. Second, *reputation systems,* which focus on eliminating spammers by constantly observing their performance, given feedback and overall satisfaction. Based on these observations, a reputation score is computed and used as a threshold for allowing or denying a worker from a particular task. The problem with such systems is that they can be exploited [5], and they suffer from both the cold start problem [6] and the challenge of computing robust, yet reliable aggregated scores. Third, *Majority voting,* which are denoted as the front-runner of the aggregation methods, where workers can be eliminated based on their deviation from the general consensus. However, it inherently incurs more costs and has its limitations, especially when the percentage of spammers in the workforce is high, see e.g. [7].

In acknowledgment of impact sourcing, many crowdsourcing platforms have been founded e.g. *RuralShores*[1]. Moreover, some so-

WebSci '16, May 22-25, 2016, Hannover, Germany
ACM 978-1-4503-4208-7/16/05.
http://dx.doi.org/10.1145/2908131.2908187

[1] http://ruralshores.com/about.html

cially-responsible quality control measures were developed to identify biased or low-skilled honest workers, such as: an algorithm separating unrecoverable error rates from recoverable bias [8], and our work on adaptive gold questions [9], and on mining irregular workers' answer patters to identify fraud [10].

3. MOTIVATIONAL USE CASES

Next, we investigate how the current quality control measures behave, with respect to honest low-skilled workers within both non-skewed and skewed crowdsourcing tasks.

3.1 Non-skewed Crowdsourcing Tasks

We refer back to our early work in [10], where we conducted a small-scale experiment, with a total of 18 volunteers. Given a set of 20 multiple choice questions, the volunteers were asked to answer the same questions twice. In the first round, they had to randomly select answers in any fashion i.e. acting as spammers. In the second round, they were asked to fairly answer the questions to the best of their knowledge i.e. acting as honest workers. The multiple choice questions were based on the verbal practice questions from the Graduate Record Examination (GRE) dataset at a medium difficulty level. Accordingly, their task was to select the right definition of a given word, without any external aid.

In figure 1, we can examine the total number of correctly answered questions for the 36 answer sets (where each volunteer answered the same set of questions twice). As to be expected, deliberate answers tend to be more correct than random answers. A closer look shows that volunteers following the random strategy had on average 40% correct answers, while those attempting to answer the questions fairly had on average 58.6% correct answers. It's important to note, that even though the dataset is in no way skewed, the random strategy indeed at times produced better overall results.

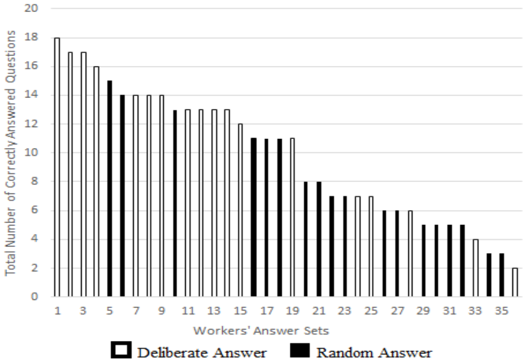

Figure 1. Deliberate versus Random Responses

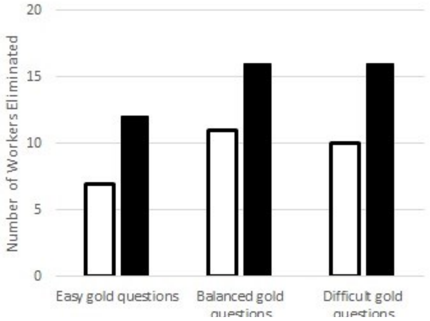

Figure 2. Number of workers eliminated for varying difficulty-levels of gold questions

Considering *reputation-based systems*, a quick look at the top ten correct answer sets reveals that the volunteer with the fifth-ranked answer set would be in fact given a higher reputation score than those with answer sets at ranks 6 to 9. That is, a spammer following a random strategy is rewarded with a higher reputation score than the honest workers. Indeed, the overall results of the spammer was better, yet solely looking at the end result, without correctly identifying the workers, leads on the long run to loses on the requestor side and unfair treatment to the workers.

On the other hand, upon using 40% easy *gold question*, with a 70% correctness level threshold, 38.8% honest workers and 66.67% spammers were eliminated. Further experiments with varying dificultly levels of gold questions (see figure 2), led to similar results; Whereas gold questions seemed to be always more inclined to penalize the spammers, still, a significant number of honest workers were also eliminated and accordingly deprived from their earnings.

3.2 Skewed Crowdsourcing Tasks

To illustrate the downfall of common quality control measures within a skewed-domain crowdsourcing task, consider the following example, which we investigated in [4].

Example: Research on quality management in Amazon Mechanical Turk (for details see [8]), set out to train an adult website classifier based on crowd-curated labels. In reality, only 15% of their web data contains adult material, with the rest 85% being suitable for general audiences. Their results illustrated that strategic spammers, who are always submitting the prevalent label, i.e. webpage doesn't contain adult content, exhibited only an error rate of 15%. In contrast, honest workers sometimes exhibited even higher error rates.

For the above example, all common measures, whether it's gold questions, majority votes, or reputation-based systems, would fail in identifying strategic spammers, because they're outwardly doing a reasonably good job. Accordingly, as soon as the skewness of the task is perceived, this easy strategy can be adopted, resulting in a seemingly good overall quality results, where in fact, it's nothing but useless.

4. CONCLUSION AND FUTURE WORK

The need for an impact-driven model is indeed undeniable. Mainly, there are two central design aspects that such a model should possess: *Quality* and *Fairness*. First, it has to ensure overall high quality for the task. Second, it has to fulfill the social responsibility of crowdsourcing, and show bias to honest workers, especially the low-skilled honest ones. This is hard, since ensuring high quality results often leads to the deployment of across-the-board discriminating countermeasures, which end up penalizing honest low-skilled workers alike. Accordingly, a clear distinction should be made between spammers and honest low-skilled workers. Moreover, in order to support low-skilled honest workers, an adaptive assignment of easier tasks would not only ensure higher quality results, but would shield the honest, yet still low-skilled workers, from definite error rates, which would reflect negatively on their reputation scores, preventing him/her from further advancement, higher paid tasks and opportunities to further develop their skills.

Our preliminary work shows that a partially redundant based technique can be employed for skewed-domain tasks, where the aggregated results of a maximum of two workers suffice and may even score higher quality than other aggregation methods like Majority Vote.

5. REFERENCES

[1] A. Sorokin and D. Forsyth, “Utility data annotation with Amazon Mechanical Turk,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, 2008.

[2] C. Lofi, J. Selke, and W.-T. Balke, “Information Extraction Meets Crowdsourcing: A Promising Couple,”in *Datenbank-Spektrum*, vol. 12, no. 1, 2012.

[3] E. Kouloumpis, T. Wilson, and J. Moore, “Twitter Sentiment Analysis: The Good the Bad and the OMG!,” in *Fifth International AAAI Conference Weblogs Soc.*, pp. 538–541, 2011.

[4] K. El Maarry, U. Güntzer, and W.-T. Balke, “A Majority of Wrongs Doesn’t Make it Right,” in *the 16th International Conference on Web Information Systems Engineering (WISE)*, 2015.

[5] B. Yu and M. P. Singh, “Detecting Deception in Reputation Management,” in *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*,pp. 73–80, 2003.

[6] M. Daltayanni, L. de Alfaro, and P. Papadimitriou, “WorkerRank: Using Employer Implicit Judgements To Infer Worker Reputation,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM),* pp. 263–272, 2015.

[7] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. W. Duin, “Limits on the majority vote accuracy in classifier fusion,” in *Pattern Analysis & Applications - PAA*, vol. 6, no. 1. pp. 22–31, 2003.

[8] P. G. Ipeirotis, F. Provost, and J. Wang, “Quality Management on Amazon Mechanical Turk,” in *Proceedings of the 2nd Human Computation Workshop (HCOMP)*, pp. 0–3, 2010.

[9] K. El Maarry, U. Güntzer, and W.-T. Balke, “Realizing Impact Sourcing by Adaptive Gold Questions: A Socially Responsible Measure for Workers’ Trustworthiness,” in *16th International Conference on Web-Age Information Management (WAIM),* 2015.

[10] K. El Maarry and W.-T. Balke, “Retaining Rough Diamonds: Towards a Fairer Elimination of Low-skilled Workers,” in *20th International Conference on Database Systems for Advanced Applications (DASFAA),* 2015.