

Crowdsourcing Data Understanding: A Case Study using Open Government Data

Yukino Baba and Hisashi Kashima
Kyoto University

Abstract

We investigate the feasibility of applying crowdsourcing to the task of data understanding, which aims to explore a set of data and discover insights from it. We consider a two-stage workflow composed of data exploration tasks and review tasks. We conducted experiments using nine datasets provided by the government of Japan and collected 114 findings and charts. The experimental results demonstrate the effectiveness of the power of crowds for data understanding and indicate the usefulness of crowd reviews for controlling the quality of crowdsourced data understanding.

Introduction

Data analysis is a labor intensive process and crowdsourcing is a promising solution to compensate for the lack of data analysts. The recent expansion of platforms for data analysis competitions (e.g., Kaggle) allows us to leverage the power of crowds for predictive modeling (Baba et al. 2014). Further, crowdsourcing is also used for data preparation including data cleaning and annotation. However, modeling and data preparation are only the two of the six data analysis phases defined by the Cross Industry Standard Process for Data Mining (Shearer 2000), and the usage of crowdsourcing in the other phases has not yet been fully explored. One such phase is *data understanding*, which aims to explore the data and discover insights. The outcomes of this phase are used for detailed investigation of the data or the development of action plans.

The crowd is believed to have the potential to contribute to the data understanding phase. There are a few studies in this line of research. Social data analysis tools such as Many Eyes (Viegas et al. 2007) have been developed, and Kaggle has launched several projects for exploring public datasets. In these studies, the contributors mainly focus on data visualization, and descriptions of any insight are not often provided. Willett, Heer, and Agrawala (2012) proposed several strategies for integrating crowdsourcing into the data understanding phase. Their strategies are based on a workflow where professional data analysts initially select subsets of the data that require further investigation, and then crowd

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

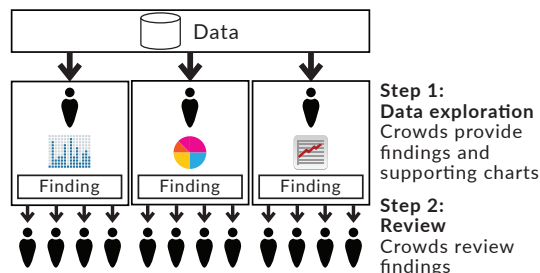


Figure 1: Workflow for crowdsourced data understanding

workers are asked to generate explanations of the data by examining the charts produced by the analysts. In their work, crowdsourcing does not cover the entire data understanding phase.

In contrast, we investigate the feasibility of applying crowdsourcing to the entire data understanding phase. Given a dataset, the data understanding phase is expected to produce reasonable findings. Figure 1 indicates our data understanding workflow incorporating crowd workers. The workflow starts with *data exploration tasks*, in which workers are asked to explore the datasets and to provide findings of the data as well as the charts that support the findings. To ensure the quality of the findings, the workflow includes *review tasks*, which uses another set of workers to evaluate whether a finding is correctly derived from the corresponding chart.

We applied our workflow to nine datasets provided by the government of Japan and obtained 114 findings and charts from crowd workers. Our results are summarized as follows: (1) 79% of the findings were correctly derived from the supporting charts, indicating that workers generated high quality findings; (2) review tasks accurately discovered the correct findings with a recall of 1.00; and (3) 85% of the findings did not overlap, indicating that worker-generated findings were diverse. Our results demonstrate the effectiveness of the power of crowds for data understanding, and indicate the usefulness of crowd reviews for controlling the quality of crowdsourced data understanding.

Case Study of Crowdsourced Data Understanding

We selected nine datasets from an open data website organized by the government of Japan. Examples of datasets are *Census of fisheries*, *Survey of research and development* and *Report on internal migration*.

In the data exploration tasks, multiple workers were assigned to each of the nine datasets. A worker provided three findings for each dataset as well as charts supporting each finding. We provided datasets in Microsoft Excel format and the workers were instructed to generate the charts using Excel. Workers were rewarded USD 0.60 for each finding and chart pair. We collected 114 findings and charts from 14 workers.

Each review task concerned a finding and chart pair and asked workers to evaluate whether the finding was correctly derived from the corresponding chart. Charts were provided in Excel format. Each worker selected one of four options: “no chart” (the Excel file did not contain a chart), “irrelevant chart” (the chart was not relevant to the finding), “correct” (the finding was correctly derived from the chart), and “incorrect” (the finding was not derived from the chart). Ten workers were assigned to each finding and each worker was paid USD 0.15 per review. We collected 1,140 reviews from 27 workers. No workers participated in both the data exploration and the review tasks.

Results

To evaluate the correctness of the findings identified by the workers, the authors (who are all data analysis experts) reviewed all 114 findings using the same metric applied in the review tasks. We found that 79% of the findings were “correct,” 19% were “incorrect,” 2% were “no chart,” and there were no “irrelevant chart” cases. These results indicate that crowd workers can have sufficient skills to provide reasonable findings without the support of professional data analysts. We observed that the workers produced reasonable findings such as “the average number of deaths in traffic accidents over the last 15 years was the highest in December” or “the average floor area of new dwellings increased by 1996, held steady for several years, and then declined.” Some of the “incorrect” findings were caused by inaccurate wording. For instance, one worker stated that *most* groups had a particular property when in fact, just five out of eight groups had it.

We investigated the accuracy of the worker reviews by comparing the results of majority voting with the groundtruth labels. Table 1 shows the resulting confusion matrix. The workers correctly classified all the “correct” findings as “correct,” thus the recall was 1.00. The workers did not overlook any of the “correct” findings in our case study. On the other hand, 17 of the “incorrect” findings were wrongly assessed as “correct” by the workers. The results show that worker judgements are more likely to be positive than expert judgements. The workers were prone to incorrectly reviewing the findings that required precise evaluation, such as the case described above where the review workers were required to notice the incorrect usage of the

Table 1: Confusion matrix of worker review results

		Crowd		
		correct	incorrect	no chart
Expert	correct	90	0	0
	incorrect	17	5	0
	no chart	0	0	2

word “most.” At the same time, the workers noticed the findings containing obvious mistakes and classified them as “incorrect.” These results indicate the effectiveness of worker reviews for discovering reasonable findings.

We then examined the diversity of findings provided by the workers by groupings sets of findings together when they indicated the same thing. The grouping revealed that there were 97 unique findings out of 114. This emphasizes that the workers explored the datasets from various points of views. Of the 97 unique findings, 87% were provided by only one worker, 11% were provided by two workers, and 2% were provided by four workers. It was infrequent that more than two workers reported the same findings in our case study. In addition, we observed that the percentage of correct findings was 87% among the unique findings provided by only one worker, 72% among those provided by two workers, and 50% among those provided by four workers. These results imply that redundant findings are likely to be incorrect and suggest that redundancy is a key factor for controlling the quality of findings.

Conclusion

To investigate the effectiveness of applying crowdsourcing to the data understanding phase, we conducted experiments using public datasets provided by the government of Japan. We observed that the workers produced high quality findings, the review tasks accurately discovered correct findings with a recall of 1.00, and most of the worker generated findings did not overlap others. Our next step is to incorporate statistical quality control methods (Baba and Kashima 2013) to assess a worker’s ability to complete data exploration tasks and review tasks in order to accurately discover reasonable findings.

References

- Baba, Y., and Kashima, H. 2013. Statistical quality estimation for general crowdsourcing tasks. In *Proc. SIGKDD*.
- Baba, Y.; Nori, N.; Saito, S.; and Kashima, H. 2014. Crowdsourced data analytics: A case study of a predictive modeling competition. In *Proc. DSAA*.
- Shearer, C. 2000. The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing* 5(4).
- Viegas, F. B.; Wattenberg, M.; van Ham, F.; Kriss, J.; and McKeon, M. 2007. Manyeyes: A site for visualization at internet scale. *IEEE Trans. Vis. Comput. Graphics* 13(6).
- Willett, W.; Heer, J.; and Agrawala, M. 2012. Strategies for crowdsourcing social data analysis. In *Proc. CHI*.