Eric Wolfe
erwolfe@siue.edu

Quality of crowdsourcing data

Introduction

Sharing information or data has been around for longer than the United States has been a country. Certain data has allowed us to go further in science then we ever thought possible. It has also won wars, saved lives, and changed how we gather data. Data is what drives our current lives. Whether it be data in a spreadsheet to calculate students' scores or data on a server that populates your Facebook feed. How does this data get collected though? Most of the time, it is inputted by users who are using that software or database. For the example of entering student scores into excel, the information comes from exams, quizzes, and homework. What about data that the community provides? Currently, companies are using the community to provide data about news events, pictures of restaurants, and traffic congestion. This type of data collection is called crowdsourcing and it's been around a lot longer then what we think. Back in World War II, allied spies were used to collect data on enemy troop movements which would be sent back to headquarters so the commanders could design a plan of attack. Now this information wasn't always correct which caused lives to be lost. This begged the question of how can we make sure the data we receive is accurate? This question is still asked to this day and many have looked into how to improve the data we receive. In this paper, I will discuss how crowdsourcing is being researched today, how it's being improved on, and what hasn't worked so well.

Problem Statement

Crowdsourcing is a great way to gather data from the community. As discussed before, it has been used for many years and it does provide a way to gather data that would not be known

otherwise. The problem that has been prevalent since the beginning of crowdsourcing which is what is the quality of the data that is being collected. Just because a spy says that the enemy is heading west, does that mean they are going in that direction or was it to avoid a broken road and now they are heading south. Should there be standards before data is considered accurate by the community or in this case spies? Another problem is what about the data we have collected in the past. What if an individual takes a picture of a restaurant in the city? At some point, you must ask if the restaurant still looks the same or has the restaurant moved to another part of the city. Should past information be revisited in the present to verify its quality? A finally problem that arises from crowdsourcing data is how does the holder of the data protect it from being tempered with. Should the community have access to delete or update information or can they only provide new data and view the data? These problems have shown up when it comes to crowdsourcing data. In this paper, I will present answers to these questions and discuss what the academic community has done to research these questions.

Related Work

Paper 1

- Crowdsourcing Translation: Professional Quality from Non-Professionals
- Translating words from one language to another
- Introducing quality control to improve crowdsourcing results
- Analyzing good and bad translations from crowd sourcing
- Cheaper to use crowdsourcing

Paper 2

- Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria

- Improving data quality from crowdsourcing annotations
- Lower costs for crowdsourcing annotations
- Found out how to sort through bad annotators
- Using quality control to find more accurate classification models

Paper 3

- In Search of Quality in Crowdsourcing for Search Engine Evaluation
- Does paying for crowdsourcing improve data quality
- Does paying more money improve data quality
- Does a individuals qualifications improve higher quality labels
- Higher pay is linked to better results
- More qualified workers improve data quality
- Crowdsourcing with lower amounts of money means that more spam data comes through

Paper 4

- Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing
- Crowdsourcing non-profit with a mission to lift people out of poverty through training and provision of digital work.
- Is a general-purpose crowdsourcing platform with built-in quality assurance.
- Experiment 1 achieved a 99% accuracy gold yield.
- Experiment 2 programmatic gold resulted in overall accuracy of 92.2% which is higher than the baseline of 85%

Project 5

- Quality Control in Crowdsourcing Systems
- Quality control approaches
- Identify open issues
- Future research

Paper 6

- Quantification of YouTube QoE via Crowdsourcing
- Assessing and modeling Quality of Experience for online video services that are based on TCP-streaming
- Using crowdsourcing to conduct user experiments
- Stalling effects QoE
- Crowdsourcing was demonstrated to be a good method for conducting QoE for online video services.

Analysis

- Papers 1, 2, 3, and 4 all have a component of how much can be saved by using crowdsourcing data instead of professional.
- Papers 1, 2, 3, and 5 look at how to improve crowdsourcing results.
- This area of research is very new but the idea of crowdsourcing data has been going on for a while.
- Questions: Does providing money to individuals who crowdsource data become an issue to see how good the quality of the data is? What kind of constraints should be generalized before using crowdsourced data? Are the individuals who provide crowdsourced data even qualified enough to provide that information?

Future Directions

- Look into rewarding individuals that provide data to you for crowdsourcing. If the reward is “better”, does that improve the results. This would be different then just offering cash. Maybe benefits to members of a subscription service.
- Does providing more constraints on crowdsourced data improve the quality of the information or limit the amount of people contributing.
- Observing if crowdsourcing will provide lower costs in all different times of fields or just some of them.
- I think in 5 years the industry will start turning more to crowdsourced data because it is becoming harder and harder to get questions answer like in the QoE paper.
- I would perform another study like the QoE but instead do it on social media.

Core Papers

- Crowdsourcing Translation: Professional Quality from Non-Professionals
- Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria
- In Search of Quality in Crowdsourcing for Search Engine Evaluation
- Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing
- Quality Control in Crowdsourcing Systems
- Quantification of YouTube QoE via Crowdsourcing

Bibliography

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E., & Dustdar, S. (2013). Quality control in crowdsourcing systems: Issues and directions. IEEE Internet Computing, 17(2), 76-81.

Hoßfeld, T., Seufert, M., Hirth, M., Zinner, T., Tran-Gia, P., & Schatz, R. (2011, December). Quantification of YouTube QoE via crowdsourcing. In Multimedia (ISM), 2011 IEEE International Symposium on (pp. 494-499). IEEE.

Hsueh, P. Y., Melville, P., & Sindhwani, V. (2009, June). Data quality from crowdsourcing: a study of annotation selection criteria. In Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing (pp. 27-35). Association for Computational Linguistics.

Kazai, G. (2011, April). In search of quality in crowdsourcing for search engine evaluation. In European Conference on Information Retrieval (pp. 165-176). Springer Berlin Heidelberg.

Oleson, D., Sorokin, A., Laughlin, G. P., Hester, V., Le, J., & Biewald, L. (2011). Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. Human computation, 11(11).

Zaidan, O. F., & Callison-Burch, C. (2011, June). Crowdsourcing translation: Professional quality from non-professionals. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 1220-1229). Association for Computational Linguistics.