Eric Wolfe
erwolfe@siue.edu

Quality of crowdsourcing data

Abstract

This paper will be analyzing several parts of quality control regarding crowdsourcing. First, I will be introducing the topic of crowdsourcing and how it has been used in the past. Next, I will look into the related work from other research groups. These research groups have several research topics in mind. Some consider the quality control of crowdsourcing that are specific to their problem. Other's consider the quality control of crowdsourcing but at a big picture level. Following that section, I analysis the papers and draw some conclusions from the papers. These conclusions build upon the papers subjects but are analyzed at a level of why it makes sense. Also, if we combine some of the topics, crowdsourced data quality can be improved be a large amount because of the restrictions that are placed on individuals. Finally, I discuss the future directions that this specific field of research will most likely go and why I believe it's important.

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

The first paper that went along with my topic was called "Crowdsourcing Translation: Professional Quality from Non-Professionals". This paper considered translating text from one language to another by using crowdsourcing. Crowdsourcing was used to help reduce the cost of translations because professional translators would charge money for each word that was translated. Another reason the authors were considering alternatives was because the money that would be needed to make a collection of texts for translations would cost a large sum of money as well. Researchers of this paper considered a way to check the quality of the translated text. They created a way to quantify the results using several different measures such as sentence-level. Using these measures, the authors could compare the crowdsourced translations against reference data. After collecting this information, they could conclude that using quality restrictions resulted in better translations from the crowdsourced translators. Better translations were not the only benefit, they also found that by using the crowdsourced translations was linked to lower costs because they did not have to pay professionals.

"Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria" was the second paper that will be used as support for this topic. This paper shared a lot of the same topics that were being analyzed in the first paper. Improving data quality was also being evaluated in this paper but instead of translations, the authors were considering annotations. That was not the only topic that was the same however, lower costs for crowdsourcing was also being considered in this paper. In the conclusion, the authors did not mention anything about the cost of using crowdsourcing. Instead, the authors discovered two important things when the research

concluded. The first was that they did figure out how to sort through bad annotators and good annotators. Secondly, the research gave the authors some insight into how to find more accurate classification models.

Following the first two papers was another paper that looked at the quality of crowdsourced data but this time the crowdsourced individuals were paid for their data. This paper is called "In Search of Quality in Crowdsourcing for Search Engine Evaluation" and focused on three important topics. Firstly, the authors considered whether paying for crowdsourced data would improve the quality of the information. Next, they considered if the amount of money they paid the crowdsourced individuals would improve the data quality or if it would stay the same. Lastly, this paper observed if a individuals qualifications improve the data quality or if it was negligible. Unlike the results of the previous paper, this paper came back with answers for each one of the topics they considered. The authors observed that when they paid crowdsourced individuals lower amounts of money, more spam would be collected. However, if the individuals were paid larger sums of money, they could link the higher pay to better results. Lastly, when individuals with better qualifications were used, the data quality would be better than non-qualified individuals.

Next, the paper titled "Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing" considered data annotation just like the second paper. Unlike the other papers being considered as related work, this paper focused on getting people out of poverty. The authors wanted to use digital crowdsourced work to lift people out of poverty. This was not the only important point of this paper. Another important topic was if a general-purpose crowdsourcing platform with built-in quality assurance could even return good quality data. Taking these research ideas in mind, the authors came up with two experiments they wanted to

consider. Firstly, they compared two effects, one was the programmatic gold and the second was manual gold. Following the first experiment was another experiment that focused on scaling gold units from a small amount to a larger amount. Each experiment was a success in the authors eyes. Experiment 1 achieve a 99% accuracy gold yield and Experiment 2 resulted in an overall accuracy of 92.2% which is higher than the 85% baseline with the programmatic gold.

Additionally, the paper, "Quality Control in Crowdsourcing Systems" is also being used in this paper because it identifies open issues and quality control approaches. This paper identifies two types of quality in crowdsourcing systems, worker's profile and task design. The authors then talk about the different approaches that can be taken when it comes to quality control. Although this paper is considered a research paper, this doesn't have any experiments or conclusions about the topic. It is more of an informational tool for individuals or companies interested in the topic.

Finally, the last paper that is being used is called "Quantification of YouTube QoE via Crowdsourcing". In this paper, the authors considering the quality of experience from online video services that are based on TCP-streaming. At the beginning of the paper, the authors explain the differences between the different streaming types and which one Youtube uses. During the research, they looked at using crowdsourcing to conduct the user experiments that went along with the research. Part of the research was addressing the issues of stall times on Youtube. If the video stalled for too long, what would the user do and why. In conclusion, the researchers did find that crowdsourcing was demonstrated to be a good method for conducting quality of experience tests for online video services.

Analysis

The first analysis that I want to analyze is the concept of saving money by using crowdsourcing. Not every paper touched on the concept but 1 of them did. This paper wanted to collect some sort of data but the authors did not want to spend large sums of money to do so. Why is money so important for collecting data? Well some of the authors touched upon this in their papers. Omar and Chris said in their paper that "Germann (2001) estimated the cost of hiring professional translators to create a Tamil-English corpus at $0.36/word. At that rate, translating enough data to even a small parallel corpus like the LDC's 1.5 million word Urdu-English corpus would exceed half a million dollars [6]." Half a million dollars is a lot of money for collecting data to make translations faster and more accurate. The Omar and Chris also said in their paper that the amount that would be spent for translations, edited translations, and rank labels would come out to $1,500 by using crowdsourcing [6]. That amount is far less than the cost to create a parallel corpus. Now this paper might have been looking at a special case but the benefits of using crowdsourcing are there.

Following that analysis, let's look at the amount of pay for crowdsourcing and if that will affect the quality of the data and if the more qualified someone is the better the quality of the data. Gabriella said in her paper that "we experimented with two pay levels, paying $0.10 or $0.25 per HIT [4]." She was considering a different pay level because the quality could be higher depending on how much you pay a person. Her results did should that this was case. Results showed that a larger pay would improve the quality of work while lower pay would increase spam and unusable labels [4]. Why is this the case though? Considering that when people are incentivized to do a good job then they will do it. This comes from rewarding someone for their work. If a person was give two choices with the first being a smaller paycheck or a larger paycheck for doing the exact same job then of course they will take the larger paycheck. This is

simple psychology. Let's look at a person's qualifications and if that will increase or decrease the quality of the data. Gabriella also questioned this in her paper when she concluded that "more qualified workers produce better quality work [4]". She said that there were side factors that could have caused this result but that is only a possibility. The important thing here is that there is a link between an individual's qualification and the quality of the data. Why are both factors being analyzed together though? Qualification and pay can be analyzed in more of a common-sense approach as well. Experience can be the difference between a good quality product and a bad quality product. On top of that, if an individual has a lot of experience and a large paycheck then they will be incentivized and they will want to do a good job to keep that paycheck. When it comes to crowdsourcing, let's observe two individuals, one with experience in being a restaurant critic and a random individual off the streets. The incentive for submitting a review is a free meal at the restaurant. Who would Google want to submit a review for a restaurant on Google Maps? Of course, they would want the critic. The critic has more experience than the random individual and the incentive is the same. This paper solidified the idea that incentives and experience are both needed when starting to collect crowdsourced data.

Previously, we have talked about reducing the cost of crowdsourcing data, giving crowdsourced individuals better incentives for their work, and checking if someone's qualified to see if they produced better work. The next topic I would like to analyze is quality control for crowdsourced data. We have previously seen that better incentives and better qualified individuals are good quality control options. What if a company or research team doesn't want to pay an individual or check their qualifications? A few of the authors have come up with ways to verify if the data they are collecting is worth keeping or not. David and his team wrote "pre-screening workers is a simple strategy: set up multiple-choice questions and ban people who do

not pass the test [5].” Setting up an exam before excepting crowdsourced individuals is a great way of doing quality control. This method touches upon the previous paper from Gabriella because as she discussed, checking someone’s qualifications is a good sign of the quality of the data. However, this instance might be a little different. Instead of checking someone’s qualifications you are simply checking whether they will be ethical and skilled in the submission of data. Next, let’s look at Mohammad and his team’s consideration into quality control. In their article, they state “we characterize quality in crowdsourcing systems along two main dimensions: worker profiles and task design [1].” First, we will check out worker profiles. Reputation and expertise is linked to a worker’s quality [1]. Having a great reputation and expertise make sense when trying to classify a worker’s quality. When an individual is famous for giving unbiased reviews for food then they should be highly considered for a food critic position. The other side of a worker’s profile is there expertise. If an individual has a master’s degree in a specific field such as computer science then they earned that degree and put a lot of work into that field. That degree means they are knowledgeable about the topic and should again be highly considered. Both have a good impact on being qualified because an individual won’t be guessing about the data they are submitting. They would have a level of understanding about what the collectors of the data are looking for. Next, task design is another part of the quality control that they looked at. Mohammad and his team also talked about the task design and had this to say, “we identify four important factors that contribute to quality about this dimension: task definition, user interface, granularity, and compensation policy [1].” First, the task definition has an important role in quality control because if the task is not clearly defined then how can you get good data. Let’s say a company wants to collect the phone numbers of all 18 to 25-year-old males but the task says to collect the phone numbers of individuals. The definition of the task is not clearly

defined here and would result in bad data. User interface is a very simple case when it comes to crowdsourcing. If a UI is hard to work with then people won't want to use it. This will reduce the amount of data you can collect or it might be bad quality data. Next, granularity gets a little more complex when it comes to a higher quality of data. There are two times of tasks, simple tasks and complex tasks. Simple tasks should always require little effort but yield good quality results. Complex tasks should be a set of simple tasks that might take longer to do. Simple tasks should be given to individuals that are less qualified and the opposite for be done for the complex tasks. Lastly, incentives and compensation policy was the last topic they considered. Stated previously, individuals respond well to higher amounts of compensation but any compensation is better than none. Giving compensation will improve the data quality and it's a good quality control topic. In this analysis, we looked at all the ways to improve data quality by putting measures into place that should prevent bad or poor data from being collected.

Future Directions

To begin, there are a few directions that data quality in crowdsourcing could go. The main direction that I believe data quality will go is researching feedback for products and services. Right now, places like Best Buy are using a verified system when submitting reviews. I believe more companies will take this approach because crowdsourced reviews for products do get out of hand when it comes to spam. In this way, companies don't need an intensive screening before starting to get reviews for products. Services might be a little bit harder to get reviews for. We have seen in the past that some parts of our entertainment or food industries have critics that are associated with either of those sectors. To be able to grow the reviews for movies, TV, and the food industries, companies can develop and research crowdsourcing solutions. I believe it will be interesting to give incentives to people who review restaurants by giving them a free

meal. This will incentives individuals and beyond that after you have reviewed more at least 10 smaller restaurants then that individuals can review larger restaurants. In this way, you can verify that the reviewer has some experience. A team could research this idea by reviewing the submitted review and analyzing it against the older style of not providing an incentive or having a quota of previous reviews. In this way restaurants could get better feedback for dishes, service, and management.

The last topic I believe will take a new direction for crowdsourced data quality is dividing up problems to lots of individuals. The constraints this research is being ethical and putting a small amount of time to solve the problem. Let's take the idea of doing a man hunt but instead of the police doing most of the work, let the community help more. Everyone can be given a video from a surveillance camera and write down if they see anything suspicious. This way the work can be divided up between a larger group and the person could be caught faster. The point of this research would be to see if you can get good quality data by using crowdsourcing to solve an issue that everyone has a common goal in.

Conclusion

In this paper, we analyzed several papers and considered why they are important for the crowdsourced data quality field of research. I introduced why crowdsourcing has be going on for a while but has not be researched heavily. After that, I introduced some of the papers that might be used in my paper and why they are relevant. Next, I concluded that providing an incentive will increase the quality of data. Also, checking an individual's expertise or qualifications will also help with the quality. That wasn't all that I observed though, I also found that crowdsourced data can save a lot of money in some fields of research or usage.

Finally, I discussed the future directions that this research might be heading in and why I believe that is the case.

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

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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).

[6] 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.