Managing Crowdsourced Human Computation

Panagiotis G. Ipeirotis
New York University
panos@stern.nyu.edu

Praveen K. Paritosh
Google, Inc.
pkp@google.com

ABSTRACT

The proposed tutorial covers an emerging topic of wide interest: Crowdsourcing. Specifically, we cover areas of crowdsourcing related to managing structured and unstructured data in a web-related content. Many researchers and practitioners today see the great opportunity that becomes available through easily-available crowdsourcing platforms. However, most newcomers face the same questions: How can we manage the (noisy) crowds to generate high quality output? How to estimate the quality of the contributors? How can we best structure the tasks? How can we get results in small amounts of time and minimizing the necessary resources? How to setup the incentives? How should such crowdsourcing markets be setup?

Their presented material will cover topics from a variety of fields, including computer science, statistics, economics, and psychology. Furthermore, the material will include real-life examples and case studies from years of experience in running and managing crowdsourcing applications in business settings.

The tutorial presenters have an extensive academic and systems building experience and will provide the audience with data sets that can be used for hands-on tasks.

Keywords

crowdsourcing mechanical turk workflow control quality assurance incentives reputation market design human computation

1. INTRODUCTION

Crowdsourcing and human computation is a relatively new research area that studies how to build intelligent systems that involve a combination of computers and humans collaborating seamlessly over the Internet. Over the last few years, due to the appearance of crowdsourcing marketplaces (e.g., Amazon Mechanical Turk, oDesk) and due to the emergence of ingenious applications (e.g., the ESP Game), we started seeing the enablement of applications that were not possible before. Typically, such applications involve the use of humans to perform some form of computation (e.g., image classification, translation, protein folding) that leverage human intelligence but challenges even the most sophisticated AI algorithms that exist today.

There are various genres of human computation applications available today. Games with a purpose (e.g., the ESP Game) specifically target online gamers who, in the process of playing an enjoyable game, generate useful data (e.g., image tags). Crowdsourcing marketplaces (e.g. Amazon Mechanical Turk) are human computation applications that coordinate workers to perform tasks in exchange for monetary rewards. In identity verification tasks, users need to perform some computation in order to access some online content; one example of such a human computation application is reCAPTCHA, which leverages millions of users who solve CAPTCHAs every day to correct words in books that optical character recognition (OCR) programs fail to recognize with certainty.

Despite the variety of crowdsourcing and human computation applications, the knowledge on how to best organize humans to work seamlessly together with machines is still in its infancy and not well understood. With the proposed tutorial, we plan to cover the recent literature in computer science that focuses on the topic of crowdsourcing, and also provide pointers to areas of research in statistics, control theory, economics, psychology, and epidemiology that provide solutions for some of the problems that we face today in crowdsourcing applications.

2. PRESENTERS

Contact Info for Panos Ipeirotis:

- Affiliation: Department of Information, Operation, and Management Sciences, Stern School of Business, New York University
- Phone: +1 (212) 998-0803
- Email: panosstern.nyu.edu
- Address: 44 West Fourth Street, New York, NY 10012 USA

Bio: Panos Ipeirotis is an Associate Professor at the Department of Information, Operations, and Management

Sciences at Leonard N. Stern School of Business of New York University. His recent research interests focus on crowdsourcing and on the use of Mechanical Turk for a variety of data-oriented applications. His work on the how to use the (noisy) information provided through crowdsourcing in order to improve data quality and data mining has received the ”Best Paper Runner Up” award at ACM KDD 2008. He co-organized two workshops on human computation and crowdsourcing (HCOMP 2009 and HCOMP 2010, both collocated with ACM KDD). He advises a significant number of firms on how to structure and manage their crowdsourced processes. He also maintains a highly-read blog, named “A Computer Science in a Business School”[1] where he blogs about research and practical topics related to crowdsourcing and Amazon Mechanical Turk.

Contact Info for Praveen Paritosh:

- Affiliation: Google, Inc.
- Phone: +1 (415) 763-7570
- Email: pkpgoogle.com
- Address: 345 Spear St, Google, San Francisco, CA 94105 USA

Bio: Praveen Paritosh is an engineer at Google where he works on systems that use humans and algorithms seamlessly. Prior to this, he helped build the Freebase, the world’s largest database of structured knowledge containing 14 million entities and 500 million relationships. A significant portion of this data was sourced by semi-automated methods with human computation in the loop. He was involved in designing and instrumenting one of the largest-scale human computation system that has collected about 3 million judgments over the course of two years. Besides building such systems, he is very interested in the social impact of crowdsourcing.

3. DURATION AND TOPICS TO BE COVERED

We propose the current tutorial to cover a 3-hour time slot (or alternative a 6 hour slot if there is enough time).

The outline below shows the topics that we plan to cover. We provide an *indicatory* (but by no means complete) set of pointers to the existing literature. Of course, the tutorial will provide pointers and discussion to a significantly wider body of material.

- **Managing quality**: One of the most important issues that everyone faces is dealing with the noise inherent in the responses provided by human users. To avoid the noise, the typical solutions are either to test users by giving them questions for which we know the answers, or by collecting redundant data and hoping that users will agree on the correct answers and disagree on the rest. In this module, we will cover traditional techniques from statistics and epidemiology that examined this topic in the past. Then, we will examine how newer techniques have been developed in computer science in the last couple of years that focus on integration of data collection with machine learning techniques, on cost-sensitive acquisition of noisy data, on estimating the quality of each user, and on how we can optimize the testing process of the users.
  – Using majority for error corrections and the limits [12]
  – Repeated labeling by focusing on uncertain cases [17]
  – Estimating worker quality [8, 10, 15]
  – Estimating difficulty of annotation [21]
  – Combining supervised and unsupervised techniques for quality assurance
  – Active testing of worker quality
- **Task design and workflow design**: Most of the techniques on managing quality rely on the assumption that the worker output is discrete or at least directly possible to evaluate as good or bad. However, increasingly many tasks are unstructured and the generated task quality is not binary but rather continuous. In such settings, it is difficult to provide the appropriate incentives for workers to generate perfect work. However, the idea of structuring human computation using iterative workflows has been proven valuable and allows us to build strong quality controls in a process that used to be hard to control. We will present results from recent research in using iterative workflows for human computation (in particular, TurkIt), and connect with results from control theory that are particularly relevant.
  – Managing free-form inputs [14]
  – Iterative workflows and control theory for quality control [11]
  – Decision-theoretic use of voting blocks [7]
- **Incentives, Game Design, and Behavioral Issues**: Even after setting up advanced quality detection schemes and after employing decision-theoretic models for controlling the workflows, there are design issues that affect the performance of the workers. For example, workers typically respond to incentives and are more willing to participate in tasks that are “fun” and are designed with an entertainment goal in mind. However, the establishment of point-based leaderboards has been shown in practice to detract users from achieving the intended goal of the task; they rather focus on optimizing the displayed score. Finally, we will discuss the well-established cognitive biases[2] for humans (e.g.,

[1] http://behind-the-enemy-lines.blogspot.com/

[2] http://en.wikipedia.org/wiki/List_of_cognitive_biases

"anchoring," the "framing effect") and present cases on how to avoid them, or even how to best take advantage of them to improve overall task execution.

  - Games with a purpose [19]
  - Leaderboards and perils of scores [9]
  - Cognitive biases, how to avoid, and how to use them [2, 3, 18]

- **Market Design Issues**: Our tutorial will also cover important aspects of economics and market design, which are important for everyone who plans on running a large-scale crowdsourcing project. For example, many crowdsourcing systems today (with Amazon Mechanical Turk being a notable example) do not allow participants to build easily an identity and a way to signal their trustworthiness to the other parties. This information asymmetry, in the presence of low quality users, quickly turns the market into a "market for lemons" in which all users (low and high quality) end up receiving low salaries. Reputation systems have been proven useful as signaling mechanisms, but over the last ten years, we also observed when they fail. We will discuss how the "micro-task" nature of the human computation challenges the existing designs schemes. Time permitting, we will discuss also lessons from design of call centers that can potentially inform the design of human computation marketplaces.

  - Information asymmetries and the market for lemons [1]
  - Reputation systems [16] and potential alternatives [4]
  - Queuing systems, call centers, and estimating time to task completion in human computation markets.

Practical Case Studies: Design Guidelines: The tutorial will place all the examples above in the context of real-life deployments. We will also have a separate section in which we will describe in detail specific applications. A special focus will be placed on Freebase, on which one of the presenters has devoted a significant amount of time and energy. Other applications that will be discussed include the ESP Game [20], Soylent [5], and the use of crowdsourcing and active learning within AdSafe Media [3] for the purposes of identifying offensive content on the Internet.

- **Lessons learnt from Freebase experience**: The ecosystem of human judges in Freebase was very different than most crowdsourcing examples: consisting of 84 paid contractors and 555 volunteers from the Freebase user community. In this section we describe examples of tasks performed, the organization and scalability aspects, and lessons learnt over the course of millions of judgments.

  - Heterogeneous tasks
  - Difficult tasks
  - Training, evaluating and feedback
  - Anonymous workers versus Employees

- **Constraints on general purpose infrastructure for human computation**: The data in Freebase [6] is collaboratively created, structured, and maintained. Automated processes form the backbone of our efforts and require human vetting to ensure that we meet our 99% accuracy standards. This quality assurance step verifies that data load processes have not compromised the integrity of a data set and will not reduce the quality of our database. From our experience, timely turn-around for these tasks is critical, empowering our data engineers to iterate rapidly and course-correct. We turn here to human computation or crowdsourcing. This section will focus on the design of the underlying infrastructure to manage millions of human judgments across a wide variety of different human computation tasks.

  - The design of RABJ
  - Constraints on a large-scale system

- **Social and Economic Impact**: There is a very large unlocked market of jobs that today require human computation. For example, it is estimated that in the US, poor data quality in healthcare alone accounts for 600 billion dollars in losses annually [13]. Opening up this marketplace will lead to creation of a large number of opportunities. In this section, we will talk about this potential impact, some examples of the social impact of such work, and challenges that lie ahead.

  - Potentially to transform by creating new jobs: "The computer is the sewing machine"
  - Examples from Samasource's work, among others
  - What TCS/Infosys did to outsourcing and connections to current trends in human computation
  - Technical challenges

4. INTENDED AUDIENCE AND BACKGROUND

The tutorial will target both researchers and practitioners that are interested in the topic but do not have any experience.

At the end of the tutorial, the practitioners will know about the best practices in structuring, running, and managing tasks on crowdsourced platforms. They will learn how to best automate the management of human contributors and how to integrate best the crowdsourced workflows with automatic machine learning techniques.

The researchers and students will be exposed to the current state-of-the-art in research methods, algorithms,

[3] http://www.adsafemedia.com/

and experiments. We will provide an overview of problems that are being currently tackled in different fields within computer science, ranging from information retrieval to computer vision. We will also connect the literature with past literature in economics, epidemiology, statistics, and psychology, allowing people to avoid reinventing the wheel but rather focus on the exciting new challenges that are unique in the crowdsourcing setting.

5. IMPORTANCE OF TUTORIAL

We consider the tutorial being important for two main reasons.

First, there is currently an explosion of research workshops in academia, all dealing with the topic of crowdsourcing and Mechanical Turk[4] While this signals the wide interest, it also indicates the wide fragmentation of the research. This tutorial is a first attempt to provide a holistic view of the area, and connect the different efforts. Also, there is a significant amount of startup companies that either rely of attempt to provide crowdsourcing services. It is important to cover best practices for them and identify areas that they need to pay attention to.

Second, the WWW 2011 is organized in India, which is a major provider of workers that participate in crowdsourcing platforms. Discussing the phenomenon and the future of crowdsourcing is of interest by itself. Given the potential of the technology to disrupt workplaces, it is of interest to *examine the current path and direction of technological advances* and see how they are expected to affect the way people work and earn their living.

6. PREVIOUS VERSIONS OF TUTORIAL

This is the first time that the tutorial will be presented.

References

[1] Akerlof, G. The market for" lemons": Quality uncertainty and the market mechanism. *The quarterly journal of economics 84*, 3 (1970), 488–500.

[2] Ariely, D. The Upside of Irrationality: The Unexpected Benefits of Defying Logic at Work and at Home, 2010.

[3] Ariely, D., and Jones, S. *Predictably irrational: The hidden forces that shape our decisions.* Harper New York, 2008.

[4] Bakos, Y., Dellarocas, C., and of Management, S. S. *Cooperation Without Enforcement?: A Comparative Analysis of Litigation and Online Reputation as Quality Assurance Mechanisms.* MIT Sloan School of Management, 2003.

[5] Bernstein, M., Little, G., Miller, R., Hartmann, B., Ackerman, M., Karger, D., Crowell, D., and Panovich, K. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology* (2010), ACM, pp. 313–322.

[6] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (2008), ACM, pp. 1247–1250.

[7] Dai, P., Mausam, and Weld, D. S. Decision-theoretic control of crowd-sourced workflows. In *AAAI* (2010).

[8] Dawid, A., and Skene, A. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics 28*, 1 (1979), 20–28.

[9] Farmer, F., and Glass, B. *Building web reputation systems.* Yahoo Press, 2010.

[10] Ipeirotis, P. G., Provost, F., and Wang, J. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (New York, NY, USA, 2010), HCOMP '10, ACM, pp. 64–67.

[11] Kern, R., Zirpins, C., and Agarwal, S. Managing quality of human-based eservices. In *Service-Oriented Computing–ICSOC 2008 Workshops* (2009), Springer, pp. 304–309.

[12] Kuncheva, L., Whitaker, C., Shipp, C., and Duin, R. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications 6*, 1 (2003), 22–31.

[13] Leitheiser, R. Data quality in health care data warehouse environments. In *System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on* (2002), IEEE, p. 10.

[14] Little, G., Chilton, L. B., Goldman, M., and Miller, R. C. Turkit: tools for iterative tasks on mechanical turk. In *KDD Workshop on Human Computation* (2009), ACM, pp. 29–30.

[15] Raykar, V., Yu, S., Zhao, L., Valadez, G., Florin, C., Bogoni, L., and Moy, L. Learning from crowds. *Journal of Machine Learning Research 11*, 7 (2010), 1297–1322.

[16] Resnick, P., and Zeckhauser, R. Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. *Advances in Applied Microeconomics: A Research Annual 11* (2002), 127–157.

[17] Sheng, V. S., Provost, F. J., and Ipeirotis, P. G. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD* (2008), pp. 614–622.

[18] Thaler, R., and Sunstein, C. *Nudge: Improving decisions about health, wealth, and happiness.* Yale Univ Pr, 2008.

[19] Von Ahn, L. Games with a purpose. *Computer 39*, 6 (2006), 92–94.

[20] Von Ahn, L., and Dabbish, L. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2004), ACM, pp. 319–326.

[4] At least 10 workshops in 2010, just within computer sciencehttp://behind-the-enemy-lines.blogspot.com/2010/10/explosion-of-crowdsourcing-workshops.html.

[21] Welinder, P., Branson, S., Belongie, S., and Perona, P. The Multidimensional Wisdom of Crowds. In *Proceedings of the NIPS 2010* (2010).