A report on the Human Computation Workshop (HComp 2009)

Panagiotis G. Ipeirotis
Stern School of Business, NYU
44 West Fourth Street
New York, NY 10012, USA
Panos@stern.nyu.edu

Raman Chandrasekar
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
RamanC@microsoft.com

Paul Bennett
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
PauBen@microsoft.com

with Edith Law, Max Chickering, Anton Mityagin, Foster Provost and Luis von Ahn

ABSTRACT

The first Human Computation Workshop (HComp2009) was held on June 28th, 2009, in Paris, France, collocated with SIGKDD 2009. This report summarizes the workshop, with details of the papers, demos and posters presented. The report also includes common themes, issues, and open questions that came up in the workshop.

Keywords

Human computation, games, game theory, labeling, crowd sourcing.

1. INTRODUCTION

The first Human Computation Workshop (HComp2009) was held on June 28th, 2009, in Paris, France, collocated with ACM SIGKDD 2009. The workshop had 33 high-quality submissions from a wide variety of perspectives. All submissions were thoroughly reviewed by the program committee and external reviewers. Given the short half-day duration of the workshop, only about a third of the submissions were accepted and may be found in the proceedings. The proceedings of the Workshop are available from the ACM Digital Library (ISBN: 978-1-60558-672-4). The workshop was well-attended with more than 40 people in the audience, pretty much filling the room. Participants selected the two best papers presented at the workshop, and these papers (listed below) will appear in *SIGKDD Explorations*:

- *Financial Incentives and the "Performance of Crowds"* by Winter Mason and Duncan J Watts.
- *KissKissBan: A Competitive Human Computation Game for Image Annotation* by Chien-Ju Ho, Tao-Hsuan Chang , Jong-Chuan Lee, Jane Yung-Jen Hsu and Kuan-Ta Chen.

The workshop web site is at http://hcomp2009.org. The workshop program is available at http://hcomp2009.org/Program.html. The proceedings of the workshop may be found at: http://portal.acm.org/citation.cfm?id=1600150&coll=ACM&dl=ACM

The workshop started with an invited talk by Luis von Ahn from Carnegie Mellon University. He described a human-powered system for translation where users pick from a selection of word translations to make coherent sentences. In the process, the user solves small translations posed as exercises. Such small translations are used to contribute to a larger translation task.

2. GAMES

The first session was on games. The first talk here described the HerdIt game, in a paper titled "User-Centered Design of a Social Game to Tag Music" by Luke Barrington, Douglas Turnbull, Damien O'Malley, and Gert Lanckriet. HerdIt uses an active learning approach to tag music. Users tag music online and then a machine learning algorithm is trained to tag a few more songs. The HerdIt game starts playing music and the players see bubbles containing tags floating on the screen (e.g., rock, pop, romantic, ballad etc). A player gets more points by selecting the bubbles that correspond to the more popular tags for the music being played. The authors have quizzes in the game (e.g., a song plays and the question "Does the singer have big hair?" appears). For the quizzes, there is a pari-mutuel prediction market running in the background, where users bet on different outcomes/answers and the winners split the common pool for the bet.

The next talk described "KissKissBan: A Competitive Human Computation Game for Image Annotation" by Chien-Ju Ho, Tao-Hsuan Chang, Jong-Chuan Lee, Jane Yung-Jen Hsu, and Kuan-Ta Chen. The authors build upon and improve the ESP game. The paper describes a technique for addressing two problems:

1. The collusion problem, where players may collude to provide bad data, and
2. The problem of limited diversity of tags, where players tend to provide easy and generic tags for an image

While taboo lists have been used for the latter, other research has shown simply presenting the players with a taboo list biases the elicited tags. The basic idea of the KissKissBan game is to convert the 2-player ESP game into a 3-player game. A new type of player is introduced, named the "blocker". The blocker enters a *hidden* taboo list to prevent the other two "matcher" players from finding a common word using the obvious words to describe the depicted image. The more words the blocker catches, the better the score of the blocker. This encourages diversity of tags, since the matcher players will attempt to provide non-obvious tags to describe the image. The blocker can also observe the words that the matchers are using, across multiple images. For example, if the two matchers develop a cheating strategy and type some unrelated words often, the blocker can catch that behavior and add such words to a blocked list. KissKissBan uses a zero-sum approach where the blocker gets the points that are being lost by the matchers. So the blocker has the incentive of entering many words that are then used by the matchers. The asymmetric pairing and the use of both competitive and collaborative mechanisms make this game novel. The area of zero-sum human computation games is relatively unexplored, and we expect to see more work around these ideas.

The third talk in this session was on the paper "Community-based Game Design: Experiments on Social Games for Commonsense Data Collection" by Yen-Ling Kuo, Kai-Yang Chiang, Cheng-Wei Chan, Jong-Chuan Lee, Rex Wang, Edward Yu-Te Shen, and Jane Yung-Jen Hsu. The authors describe the "Rapport" game to build "common-sense" ontologies using virtual pets that are being 'fed' knowledge by the player's friends (e.g. in a Facebook-like setting). The game has some quiz-like templates (e.g., X likes-to Y) which are then filled in by friends of the player using reasonable values (e.g., "a student likes-to have no homework"). To make it fun, the pets compete online playing such quizzes, and become smarter, getting "smart points". The pets get points when they give the same answers as the pets of other owners. The smartest pets that have the most knowledge and give the most sensible answers appear in the leaderboard.

3. HUMAN COMPUTATION IN PRACTICE

The next session consisted of a number of demos and posters. The demos included:

- The Phrase Detectives Game, from the University of Essex, to help create large-scale linguistically annotated corpora. http://anawiki.essex.ac.uk/
- Picture This, a game from Microsoft to elicit preference judgments to rank images for an image query. http://www.clubbing.com/Pages/Games/GameList.aspx?game=Picture_This
- Page Hunt, also from Microsoft, to garner data to improve web search relevance, going from web pages to queries. http://pagehunt.msrlivelabs.com/PlayPageHunt.aspx
- TurkIt, from MIT, providing a library and toolkit for 'iterative tasks' on Mechanical Turk.
- Seaweed, also from MIT, to help design economic games.
- Search War, from CMU, to collect, for a given web page, its relevance and salience.
- Magic Bullet, from Newcastle University, to streamline the robustness evaluation of CAPTCHAs. Thumbs-UP, from Yahoo!, to rank search results.
- Games for Games, from Est Creativity Rising and the University of Bremen, examining how human computation game design and scoring approaches affect the quality of data gathered.

There were five posters included in the program:

- In *From Active Towards InterActive Learning: Using consideration information to improve labeling correctness,* Abraham Bernstein and Jiwen Li suggest that active learning algorithms should help raters improve their performance by using 'consideration information'.
- Dorin Morrison *et al* in *TagCaptcha: Annotating images with CAPTCHAS* exploit the need for human verification to label images for keyword-based retrieval.
- Peter Faymonville *et al* in their paper *CAPTCHA-based Image Labeling on the Soylent Grid* explore usability issues in an open labeling platform built for vision researchers.
- Osamuyimen Stewart *et al* look at *Designing Crowdsourcing Community for the Enterprise* and suggest that we need to identify the right social objects and use that to design (not necessarily monetary) incentives to motivate and sustain participation in enterprise crowdsourcing.
- Trevor Burnham and Rahul Sami, in *A Reputation System for Selling Human Computation* present a model and discuss how 'partial verification' can be used to eliminate mistrust in reputation systems, and how this can help human computation become more efficient.

4. GAME THEORETIC ANALYSIS

The third session was on game theory.

The position paper by Shaili Jain and David Parkes, "The Role of Game Theory in Human Computation Systems," gave an outline of promising directions for research in this area. The basic idea is that human computation may benefit by using game theoretic concepts to improve the design of the games, just as the use of game theory solved problems (e.g. free riding) in settings like P2P networks. The presentation included a brief introduction to game theoretic analysis of some games and systems (e.g., PhotoSlap, ESP game, and Yahoo Answers). The talk advocated a modeling of user actions (e.g. in the ESP games players select "easy" or "difficult" words), the corresponding costs and benefits for the users, and how these affect the outcome of the game. The nice outcome is that game theory helps to predict the equilibrium/stable state of these games (in the ESP Game, players have the incentives to enter "easy" words). The high-level take away from the talk: it is good to build a game-theoretic model of each game, so that we can see how robust the game is to perturbations of design options.

The game theory discussion continued with the paper "On Formal Models for Social Verification" by Chien-Ju Ho and Kuan-Ta Chen. The authors describe how to use game theory to show the effect of sequential verification vs. parallel verification. Parallel verification is the process by which two users submit an answer for a question, and if it matches they get a reward. Sequential verification is where the user submits an answer that needs to match a known "correct" answer. The paper provides the corresponding equilibria that result from these mechanisms.

5. LABELING COST AND EFFICIENCY

The next session on labeling cost and efficiency started with the paper "Efficient Human Computation: the Distributed Labeling Problem" by Ran Gilad-Bachrach, Aharon Bar Hillel and Liat Ein-dor, who tackle the following problem: comparing tags introduces errors when what is desired is concept equality. That is, using humans we can collect labels and tags that describe an object (e.g., an image). When the number of possible labels is large, then we will start seeing consistency problems as different labelers use different vocabularies to create their labels. For example, labelers may provide correct answers but, due to polysemy, they may end up giving superficially different labels, even though they mean the same thing ("truck" and "lorry").

Conversely, they may give the same label even though they mean different things (e.g. "greyhound" the dog, and "greyhound" the bus company). The authors describe graph-theoretic algorithms that can be used to resolve such problems and provide bounds on the optimality of the proposed approaches. There are still questions about how to deal with the fact that the same user may not be self-consistent over time, and how to deal with users of various degrees of reliability.

The last talk of the workshop was on "Financial Incentives and the Performance of Crowds" by Winter Mason and Duncan J. Watts. The authors examine how Mechanical Turk workers (a.k.a. “Turkers”) respond to various levels of payment. The result was that Turkers respond to higher payments by doing more work, but that the quality of the work was not sensitive to the size of the payment. In a test of anchoring, the authors asked Turkers if they felt that they have been fairly compensated. The Turkers reported that they were underpaid, valuing their work 2-3 cents more per HIT compared to the actual payment. Particularly interesting was that this difference was consistent across levels of payment.

The workshop concluded with a dinner sponsored by Microsoft Research and Carnegie Mellon, where the participants continued the discussions of the day.

6. THEMES

The overall themes that emerged from the workshop were rather clear: on the one hand, there is the experimental side of human computation, where researchers are trying to devise new incentives for users to participate, new types of actions, and new modes of interaction. This contains work on new programming paradigms and game templates designed to enable rapid prototyping, allow partial completion of tasks, and aid in reusability of game design. On the other hand, we have the more abstract/theoretic side, where researchers are trying to model these actions and incentives to examine what theory predicts about these designs. Finally, there is work that examines what to do with the noisy results that are being generated by such games and systems: how can we best handle noise, identify labeler expertise, and use the generated data for data mining purposes?

We noted common themes, issues, and open questions throughout the day and synthesized those into the following list of more pressing open questions which we posed to the audience in the concluding session of the workshop.

Game Design

- What are other models of asymmetric pairing than that introduced by KissKissBan? How are these tied to zero-sum games and do they show different robustness to exploitation than the currently dominant collaborative framework?
- Is it useful to use social networks and social credit in the process of garnering data? Can such an approach be used to synthesize data for personalization rather than the global approach often taken to data mining using human computation?

Human Computation in Practice

- Can we have a declarative programming language to handle the more mundane aspects of game building?
- Sometimes a game really fits a problem better, either because of the underlying characteristics or via provable game theory results. Can we get Turkers to play a game rather than solve a HIT?
- As a research community, how can we drive traffic to games?

Game Theory and Human Computation

- What is the value of single answers from one player versus the overall target we are trying to converge on from a data mining perspective?
- Are there equilibria for [all] major classes of human computation games?
- We need game design for different classes of games - what classes do we consider?
- What is optimality in the context of human computation and how do we prove it? For example, from a data mining perspective, getting noisy labels from a human judge or labeler (a 'teacher') may be okay if the noise can be bounded or estimated in some way.
- What types of exploitation break the system irrecoverably - versus soft exploitations like single biased labelers who can be identified and post-processed out of the data during mining?

Labeling Cost and Efficiency

- What is the value of a 'teacher' here? Not everyone is equal, but how do we identify and use the better, more knowledgeable players?
- How much time should we spend on designing a game? Is it better in some cases to just get data annotated by other means than to spend time in designing games to get annotations? Is there a simple procedure to predict cost-benefit before investing too much time?
- Although most of human computation has been focused on reducing costs of data annotations, we often overlook a class of tasks where no individual can solve the problem or no individual knows all the answers. What important problems fall within this class?
- What is the number of labels vs. quality trade-off?
- What if teachers give correct answers with different probabilities? How can we model and use this?
- How do we factor in the skill of a teacher and the value of a label?
- How do we automatically use teacher expertise and label value to choose what to pay?

7. CONCLUSION

HComp2009, the first workshop on Human Computation, was extremely successful in reaching a number of disciplines and getting high quality papers from a number of people active in this area. The organizers look forward to organizing the next year’s workshop on Human Computation. The Workshop has also assembled a Wiki bibliography in the area, available at http://hcomp2009.wikispaces.com which we hope will be widely used and extended.

8. ACKNOWLEDGMENTS

The workshop organizers wish to acknowledge the sponsorship and support from Microsoft Research and Carnegie Mellon University.

The workshop organizers also specifically call out and thank Edith Law for all the energy and effort she put into the organization of the workshop.

About the authors:

Panagiotis G. Ipeirotis is an Assistant Professor at the Department of Information, Operations, and Management Sciences at Leonard N. Stern School of Business of New York University. His area of expertise is databases and information retrieval, with an emphasis on management of textual data. His research interests include web searching, text and web mining, data cleaning and data integration. He received his Ph.D. degree in Computer Science from Columbia University in 2004 and a B.Sc. degree from the Computer Engineering and Informatics Department (CEID) of the University of Patras, Greece in 1999. He is the recipient of two Microsoft Live Labs Awards, the "Best Paper" award for the IEEE ICDE 2005 conference, the "Best Paper" award for the ACM SIGMOD 2006 conference, and a recipient of a CAREER award from the National Science Foundation. (pages.stern.nyu.edu/~panos/)

Raman Chandrasekar is a Researcher in the Text Mining, Search and Navigation group in Microsoft Research, Redmond. His work at Microsoft has focused on topics at the intersection of information retrieval and natural language processing. His current interests include aspects of human computation, targeted/topical search, and search & advertising relevance. (research.microsoft.com/en-us/people/ramanc/)

Paul Bennett is a researcher in the Context, Learning & User Experience for Search (CLUES) group at Microsoft Research where he works on using machine learning technology to improve information access and retrieval. His recent research has focused on pairwise preferences, human computation, text classification, sensitivity, calibration, and combination techniques. Prior to joining MSR in 2006, He completed his dissertation on combining text classifiers using reliability indicators in 2006 at Carnegie Mellon where he was advised by Profs. Jaime Carbonell and John Lafferty. From 2005-2006, he served as the Chief Learning Architect on the RADAR project at Carnegie Mellon. (research.microsoft.com/pauben/)