Emerging Theories and Models of Human Computation Systems: A Brief Survey

Rajarshi Das and Maja Vukovic
IBM T.J. Watson Research Center
Hawthorne, NY, USA

ABSTRACT
The ubiquitous access to computing systems has spurred the design of a variety of human computation systems, each aiming to harness the power of human computation to tackle problems that cannot be solved by today's computers alone. However, the rapid pace of progress in developing increasingly complex human computation systems has far outstripped the development of abstract models to evaluate, compare and generalize such systems. In this paper we present a brief thematic overview of the nascent theories and models of human computing systems by grouping them along key design choices in building such systems.

Author Keywords
Human computation, crowdsourcing, theory and models.

ACM Classification Keywords
H.5.3 Group and Organization Interfaces: Collaborative computing, Web-based interaction.

General Terms
Design, Human Factors.

INTRODUCTION

Human computation is a rapidly growing area that aims to leverage the combined processing power of a multitude of humans to solve computational problems that computers alone cannot yet solve. A related enterprise is crowdsourcing, which aims to outsource tasks that are traditionally performed by designated human agents to an undefined large group of humans. Ingenious and successful examples of human computation and crowdsourcing systems abound in science, industry and the public sector. In science, a well known example is the Galaxy Zoo project [1] which engaged over 100,000 volunteers to classify the entire Sloan Digital Sky Survey (SDSS) catalogue of over one million galaxies, with each galaxy being viewed an average of 38 times. Enterprises are nowadays engaging online experts for a variety of tasks, ranging from classification and transcription [2], to generating new business and research ideas [3]. A striking example of such efforts involved GoldCorp [4], a gold mining company, which made available to the public geological survey data from its Red Lake, Ontario property, offering prizes of up to $575,000 to participants who could analyze the data and suggest locations where gold could be mined. In the public sector, local governments are soliciting not just data about city infrastructure and current conditions, but also ideas and feedback on its budgeting process [5]. As an example, the city of Pittsburg [6] engaged over 1600 citizens during winter snow storms to report street conditions.

In spite of the rapid pace of experimental progress in designing and running increasingly complex human computation and crowdsourcing systems, the concomitant development of analytical models has been limited in scope and complexity. Unlike the field of computer science, where the development of formal models of computation helped to shape the progress of reliable computing systems later on, there is no overarching theoretical framework to guide the myriads of human computation systems working today. With the ever-increasing number of human computing systems and their applications, there is a need for developing a principled way of conceptualizing them. Moreover, abstract models can potentially help in evaluating, comparing, and generalizing across different instances of human computation systems by factoring in (or factoring out) design choices in goal-definition, user-interface, incentive schemes, algorithms and approaches used [7]. Indeed, the development of complexity theory for systems involving human computation has been argued as one of the five deep questions facing computer science today [8].

With some of the above concerns in mind, in this paper we attempt to provide a brief survey of some of the nascent theories and models of human computations and crowdsourcing systems that have been proposed in the literature. The main thrust of our effort is not in covering all proposed models and approaches, but in providing a thematic overview of the interesting models grouped along key design choices in building human computation systems.

UbiCrowd'11, September 18, 2011, Beijing, China.

In particular we first focus on suggested frameworks for taxonomy and classification of human computation systems, which in turn suggests groupings of research on theories and models into following categories: a) Role of Incentives b) Quality management and verification schemes, and c) Evaluation frameworks. We conclude with a set of research directions distilled from our analysis.

TAXONOMY AND CLASSIFICATION

Given the diversity of human computation systems, it is of no surprise that a number of different classification or taxonomic schemes have been proposed in the literature. Employing a framework that is similar to those in the field of organizational design, Malone et al. [9] identify four key building blocks in human computing systems: (1) The Goal: What is being accomplished? (2) Process: How is it being done? (3) Staffing: Who is performing the task? (4) Incentives: Why are they doing it? The authors study existing human computing systems aiming to provide a comprehensive classification of the different schemes employed within each of the four building blocks.

While Erickson [10] also classifies crowdsourcing systems into four groups, he employs the four-quadrant model from the Computer Supported Cooperative Work (CSCW) domain to propose a scheme based on a system's distribution over time and space, and derives four categories of crowdsourcing: (1) *Global crowdsourcing* refers to applications that have crowds contributing at different times (often as a continuous effort) and from different locations. (2) co-located crowd forms *Geocentric crowdsourcing*, where crowds complete tasks related to the same location, however not necessarily at the same time. (3) *Audience-centric crowdsourcing* groups crowd contributions made at the same time. (4) *Event-centric crowdsourcing* engages crowd for a particular event that has a clearly defined duration.

Quinn and Bederson [11] propose a more in-depth classification for human computation systems along the following dimensions: motivation, quality control, aggregation, human skill, process order and task-request cardinality; with a goal to identify parallels between systems and reveal opportunities for new research. For each dimension they offer a set of sample values that they can take. For example, motivation can be provided in the form of the pay, altruism, enjoyment, reputation and implicit work. Aggregation refers to how the problem is being solved and contributions combined. Process order refers to the way different actors, workers, requestors and system, in the human computation process, are orchestrated. For example, does the system generate tasks for a worker, which in turn become part of the requestor's goal, as in CAPTCHA [12]; or if the requestor submit tasks for worker to perform, as in Amazon's Mechanical Turk (AMT), known as the marketplace model.

ROLE OF INCENTIVES

Given that the success of human computation and crowdsourcing systems hinges critically on user participation, a wide variety of incentive schemes have been adopted in encouraging user participation and in eliciting useful user behavior. DiPalatino and Vojnovic [13] focus on the relationship between user participation and incentives in a model of crowdsourcing where strategic users select among, and subsequently compete in, a collection of contests offering different rewards. They find that user participation rates can scale logarithmically with the magnitude of the offered reward and offer suggestions for structuring incentives to facilitate users in selecting their tasks.

Other researchers have looked into the relationship between financial incentives and performance by marshalling ideas and tools from economic theory and social science. Mason and Watts [14] have experimentally studied the effect of financial incentives on performance in AMT to show that increased incentives increased the quantity, but not the quality of work. Workers who were paid more believed that their work was valued more, and were no more motivated to work harder than the workers who were paid less. The authors also note that the structure of the incentives could result in better work for less pay. Horton and Chilton [15] build on the work by [14] by developing a simple rational model of labor supply in crowdsourcing systems. The model factors out the reservation wage—the key parameter in a labor supply model—by making it invariant with respect to the experimental parameters. In experiments in AMT, the authors find mixed evidence for the rational model. While workers were found to be sensitive to price incentives, they seemed to be insensitive to the magnitude of the task-completion time. The authors explain the divergence between the theoretical predictions and experimental results by showing that some of the workers followed a suboptimal strategy of aiming to reach salient target earnings instead of maximizing their total earnings.

QUALITY CONTROL AND VERIFICATION SCHEMES

While human computing systems can leverage the power of the crowd, they are often hamstrung by the fact that none of the humans in the system are aware of the ground truth and their expertise varies across the problem space. Yan et al. [16] explores challenges posed by human computing systems to the active learning scenario where ground truths are not available and where multiple labelers/annotators, with varying expertise, are available for querying. This setup, which is often adopted in human computing systems, raises the question: which data sample should be labeled next and which annotator should be queried to benefit learning model the most. The authors propose a probabilistic multi-labeler model and provide an optimization formulation that allows the selection of the most uncertain sample and the most reliable annotator (in

the model) to query the labels from for active learning. Such approaches to automatically choosing the most appropriate annotator for a particular task can make learning more scalable and efficient.

In mobile crowdsourcing, where there is increased opportunity for repeated interactions, Eagle [17] advocates sending out the same task repeatedly to verify that the same response is being received from multiple, independent users. Given a history of noisy responses from a set of error-prone individuals, expectation-maximization (EM) approaches are used to infer correct answers and estimate users' accuracy levels. Individual accuracy conditioned on task type is also be inferred and the results are used to guide future task allocations.

Many of the popular human computation systems deal with "Games with a Purpose" (GWAP) which attempts to harness useful work from humans for free in AI-hard problems. One of key issues in such GWAPs is to avoid irrelevant inputs from users by allowing users to verify answers from others. In one of the first works to provide a game theoretic analysis of a GWAP, Ho et al. [18] focus on a game for semantic annotation of images called the PhotoSlap game. They show that PhotoSlap can arrive at the subgame perfect Nash equilibrium with the target strategy when players are rational and do not engage in collusion. The authors suggest that the players can be informed of a default strategy, the target strategy, in advance to satisfy subgame perfect equilibrium. Ho and Chen [19] have applied game theory to analyze two fundamental social verification mechanisms, simultaneous verification and sequential verification, in more popular GWAPs such the ESP game. Equilibrium analysis shows that sequential verification leads to a more diverse and descriptive set of outcomes than simultaneous verification, while the latter is stronger in ensuring the correctness of verified answers.

The issue of creating high-quality content also arises in peer production systems such as online question-and-answer forums. Is there a way to allocate the available attention from viewers amongst the contributions– a mechanism– that encourages high-quality contributions, while also maintaining a high level of participation? Ghosh and Hummel [20] propose a game-theoretic framework to analyze and compare mechanisms employed in systems with user-generated content. They assume a model where the contributors are strategic and are motivated by the amount of exposure their content will receive. The contributors also have a cost of participation that increases with the quality of their content. The authors compare several mechanisms that use viewer ratings to allocate attention to content including a rank-order mechanism, where contributions are allocated positions on the page in decreasing order of their ratings, and a proportional mechanism which distributes attention in proportion to the number of positive ratings. The rank-order mechanism is more widely adopted in human computing systems, while the proportional mechanism can be viewed as a natural and more 'fair' alternative. The authors show that while the proportional mechanism always reaches equilibrium, its convergence to optimality depends on the how the number of potential contributors grows with the number of viewers. The rank-order mechanism does not suffer from such dependence, and elicit higher quality contributions in equilibrium than the proportional mechanism.

EVALUATION FRAMEWORKS

In contrast to purely theoretical models for human computation and crowdsourcing systems, other researchers have taken a quantitative and experimental approach to evaluating and understanding such systems. In building platforms that facilitate the evaluation and analysis of existing systems, they hope to distill the requirements for more usable, efficient and effective systems. Ipeirotis [21] examines and discusses a set of operational characteristics of AMT. As described earlier, AMT is crowdsourcing marketplace for human intelligence tasks (HITs), which are often of low granularity (atomic) and cannot be successfully automated, such as transcription, high quality classification, or image labeling, etc. AMT has also gained popularity in the computer science community, where it is used for user studies [22]. Completion of tasks on AMT is typically awarded a small payment (from few cents to a about a dollar). Since 2009 Ipeirotis has been collecting data on HITS to analyze the AMT properties. Iperiotis [23] reports that majority of workers on AMT are from US and India, they tend to be younger and have lower incomes and smaller families. Iperiotis also analyses top tasks requestors, and singles out aggregators offering quality assurance layer on top of AMT.

When it comes to price distribution of the tasks, Iperiotis observes that only 10% of tasks are offered at the rate of 0.02$ or less, 50% are above 0.10$, and 15% are priced at 1$ or more. As a result, Iperiotis poses a question of automated task pricing, depending on the nature of the task, existing competition, expected activity level, desired completion time and priory activity, to name a few factors. Finally, Iperiotis indicates that task completion time follows a power law, and reports on interesting patterns in task posting and completion around different week days, observing a drop in completion on Monday, as a result of lower requests during the weekend.

Heyman and Garcia-Molina [24] built Turkalytics, a scalable analytics layer for human computation systems that has been shown to handle up to 100,000 task requests per day. They present a state model for worker interaction called Search-Continue-RapidAccept-Accept-Preview (SCRAP), to capture workers behavior when searching and selecting tasks to work on, in contrast to the simple Search-Preview-Accept (SPA) model. They analyze number of user agents, IP address, cookies and number of views to identify unique workers (and their location), as some may register with misleading demographic information. Whilst they

found it rare for workers to use multiple profiles, they did detect one user with seven different credentials. They verify Iperiotis' findings that most workers come from USA and India. They suggest the form of payment provided for these countries, micropayments, comes naturally to the users.

RESEARCH DIRECTIONS

The potential application of human computing systems on a broad spectrum of tasks derives from the richness of the computational capabilities of humans. It is the very richness and complexity of an individual human processing element that makes it difficult to analyze or predict the emergent behavior arising from the interactions of a multitude of humans. This further exacerbates the problem of achieving desired global behavior by designing incentive schemes for coordinated problem solving among individual human agents who follow their own goals. With increasing system complexity, the field of human computation system matures will need to draw upon theories and models from a number of fields including AI, multi-agent systems, game theory, operations research, economics, social sciences, and human-computer interaction. As an example, Jain and Parkes [25] envision how game theory can play role in designing and understanding of human computation systems. Developing a comprehensive framework for human computation that includes human-centric measures such as costs, availability, dependability or usability will remain as a key goal for the future.

REFERENCES

1. Clery Daniel. Galaxy Zoo Volunteers Share Pain and Glory of Research. Science, Vol. 333, No. 6039. (8 July 2011), pp. 173-175.
2. www.crowdflower.com
3. www.innocentive.com
4. GoldCorp Challenge. http://www.goldcorpchallenge.com/
5. http://oakgov.ideascale.com/
6. How's my street? Crowdsourcing app lets you know. IT World. February 2010.
7. Kulkarni, A. The Complexity of Crowdsourcing: Theoretical Problmes in Human Computation. CHI-11 Workshop on Crowdsourcing and Human Computation.
8. Wing. J. Five Deep Questions in Computing. Communications of the ACM. Vol. 51. Number 1. 58-60. 2008.
9. Malone, T., Laubacher R., and Dellarocas C. Harnessing Crowds: Mapping the Genome of Collective Intelligence. Working Paper No. 2009-001. MIT Center for Collective Intelligence. 2009.
10. Erickson, T. Some Thoughts on a Framework for Crowdsourcing. Workshop on Crowdsourcing and Human Computation. 2011.
11. Quinn, A. J., Bederson, B. B. Human Computation: A Survey and Taxonomy of a Growing Field. In Proceedings of the 2011 annual conference on Human factors in computing systems (CHI). 2011.
12. von Ahn, L., Blum, M., Hopper, N., and Langford, J. CAPTCHA: Using hard AI problems for security. In Proceedings of Eurocrypt, 2003.
13. Palatino D. and Vojnovic M. Crowdsourcing and All-pay Auctions. Proceedings of the 10th ACM conference on Electronic Commerce. 2009.
14. Mason, W. and Watts, D. Financial Incentives and the "Performance of Crowds". Proceedings of ACK SIGKDD Workshop on Human Computation (HCOMP), 2009.
15. Horton, J. and Chilton L. The Labor Economics of Paid Crowdsourcing. Proceedings of the 11th ACM conference on Electronic commerce. 2010.
16. Yan Y., Rosales, R., Fung, G., and Dy. J. Active Learning from Crowds. Proceedings of the Int. Conf. on Machine Learning (ICML), 2011.
17. Eagle, N. txteagle: Mobile Crowdsourcing. Internationalization, Design and Global Development, LNCS 5623, 447-456. 2009.
18. Ho, C., Chang, T., and Hsu J. PhotoSlap: A Multi-player Online Game for Semantic Annotation. Processings of the 22nd AAAI Conference/ 2007
19. Ho, C. and Chen K. On Formal models of Social Verification. Human Computation Systems. Human Computation Workshop (KDD-HCOMP'09). 2009.
20. Ghosh, A. and Hummel, P. A Game-Theoretic Analysis of Rank-Order Mechanisms for User-Generated Content. Proceedings of the 12th ACM Conference on Electronic Commerce. 2011.
21. Ipeirotis, Panagiotis G., Analyzing the Amazon Mechanical Turk marketplace. XRDS 17, 2. 2010.
22. Kittur A., Chi E., and Suh B. Crowdsourcing User Studies With Mechanical Turk. In Computer Human Interaction, pages 453–456. ACM, 2008.
23. Ipeirotis, Panagiotis G., Demographics of Mechanical Turk (March 2010). NYU Working Paper No. ;CEDER-10-01. Available at SSRN: http://ssrn.com/abstract=1585030
24. Heymann Paul and Garcia-Molina Herman. Turkalytics: analytics for human computation. In Proceedings of the 20th international conference on World Wide Web (WWW '11). 477-486. 2011.
25. Jain, S. and Parkes, D. The Role of game Theory in Human Computation Systems. Human Computation Workshop (KDD-HCOMP'09). 2009.