Human Computation: Experience and Thoughts

Kuan-Ta Chen
Institute of Information Science, Academia Sinica
ktchen@iis.sinica.edu.tw

In this position paper, I will first summarize our previous work on human computation, and briefly discuss our ongoing work which we consider potential and interesting. Finally, I will elaborate my thoughts on how to further extend this field for deeper and broader use.

EXPERIENCES

Performance Analysis of GWAP

We started by studying the properties of GWAP (Games With A Purpose) systems and identified the need for proper puzzle selection strategies in order to collect human intelligence in an efficient manner. Based on our analysis, we have proposed the Optimal Puzzle Selection Strategy (OPSA) [3, 7] to improve the efficiency of GWAP systems. Using a comprehensive set of simulations, we demonstrated that the proposed OPSA approach can effectively improve the system gain of GWAP systems, given that the number of puzzles in the system is sufficiently large.

Game Theoretical Modeling of GWAP

In GWAP systems, users are normally required to input answers for questions proposed by the system, e.g., descriptions about a picture or a song. Since users may bring up irrelevant inputs intentionally or carelessly, and often the system does not have "correct" answers, we have to rely on the users to verify answers from others. We call this kind of mutual verification of users' answers "social verification."

In [6], we have proposed formal models for two fundamental social verification mechanisms, namely, *simultaneous verification* and *sequential verification*. By adopting a game-theoretic approach, we perform an equilibrium analysis which explains the effect of each verification mechanism on a system's outcome. Our analysis results show that *sequential verification leads to a more diverse and descriptive set of outcomes than simultaneous verification, though the latter is stronger in ensuring the correctness of verified answers*. Our experiments on Amazon Mechanical Turk, which asked users to input textual terms related to a word, confirmed our analysis results.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

Competitive Elements in GWAP

GWAP systems are generally susceptible to the problems of *cheating* and *homogeneous inputs*. To address both issues, we have proposed the human computation game, KissKissBan (KKB), for image annotation [5]. KKB is unique in that it is the first to integrate both collaborative and competitive elements in the game design, which enables two nice properties over earlier human computation games. Firstly, since the blocker is encouraged to detect and prevent coalition between the couples, these efforts naturally form a *player-level cheating-proof mechanism*. Secondly, to evade the restrictions set by the blocker, the couples would endeavor to bring up a more diverse set of image annotations. The results of experiments on Amazon Mechanical Turk and a gameplay survey involving 17 participants showed that KKB is a fun and efficient game for collecting *diverse* image annotations.

Cheat Detection for Crowdsourced Ratings

Rating is commonly used to measure the quality of multimedia content and systems. However, conducting rating experiments is inefficient for three reasons: 1) It is difficult to rate content quality by giving an absolute score; 2) it is not economic in terms of experiment time and cost; and 3) the experiments need to be conducted under the constraints of working hours, computer hardware, and place.

In our earlier work [2], we address these issues by taking the MOS (Mean Opinion Score) methodology as a starting point. Traditionally, multimedia/CHI researchers conduct MOS experiments to quantify user's satisfaction with multimedia content and systems, that is, experiment participants are asked to rate certain stimulus with an integer score from 1 to 5. The average score is taken as the quality of the content. We propose a novel experiment framework to solve all the three aforementioned problems. By using *paired comparison* and *an algorithm to verify the consistency of each participant's input*, the proposed framework reduces the difficulty of obtaining quality judgments, and enables researchers to invite Internet users to participate in their quality assessment experiments. The evaluation results demonstrate the proposed framework enables researchers to crowdsource their experiments to an Internet crowd *without risking the quality of the results*, and, at the same time, *obtain a higher level of participant diversity at a lower monetary cost*. We have also introduced its implementation as an open platform to general readers in [1].

WORKS IN PROGRESS

Quality Assurance Framework

Based on our proposed anti-cheating framework for rating-based user studies [2, 1], we are further developing a generalized framework for assessing and assuring the quality of user inputs from crowdsourcing tasks.

Digital Archiving

Having participated in the TELDAP (Taiwan e-Learning and Digital Archives Program) project[1] for longer than three years, we have identified that a number of key techniques in digital archiving can be largely enhanced by using human computation and/or crowdsourcing. For example, one of the key missions of TELDAP is to digitalize ancient Chinese documents. In the past 12 years, TELDAP spends millions of US dollars every year to have human verifiers manually verify, one word by one word, the characters recognized by OCR software. The progression is extremely slow because the speed of human verification is quite limited (approx. 6 million Chinese characters per human-year), and, to meet the accuracy standard (i.e., an error rate lower than 1 in 20,000 characters) of TELDAP, each document needs to be sequentially verified by at least three separate verifiers.

The reasons that existing crowdsourcing solutions like reCAPTCHA [12] are not used include 1) reCAPTCHA requires that each word has been properly cropped, but the layout of many ancient documents is blurred and complicated and cannot be well analyzed by OCR software, and 2) many characters in ancient documents cannot be displayed and typed in modern operation systems because they are not included in the any of the standardized charset, including Unicode.

Because of the huge volume of ancient Chinese documents which remains un-digitized, TELDAP will continue this effort in the foreseeable future. Also, we observe that the situation we are facing is actually more general than one would initially think; for example, automatic recognizing text in comic drawings and in hand-written documents is still far beyond the capability of current OCR software. Therefore, we are putting efforts into developing mechanisms to enable the crowd help in segmenting complex documents, recognizing text, and verifying recognition outcomes.

Emergency Informatics

When a disaster occurs, social microblogging services (e.g., Twitter and Plurk) and SMS (short message services) have been proved valuable in that they can help gather real-time updates from social reporters (some of them may be witnesses of crisis) who are at vantage points and able to access first-hand and trustworthy information [11, 10, 9]. Such crowdsourcing model manifested its usefulness and unprecedented role in recent Haiti earthquake response. However, the information diversity of crowdsourcing inevitably introduces certain degree of *information non-verifiability*. If a social report claimed that a crisis just occurred or an emergent need of human forces and goods somewhere, how should a crisis response team react? It would be another crisis if authorities and disaster relief units respond to every social report without information verification. However, to verify the trustworthiness of social reports remains an open problem since traditional methods for information quality assurance, such as reputation management, may not be applicable in diaster response scenarios.

Therefore, before fully embracing the power of mass crowd, we have to be careful and aware of the risks from false information. We are developing a framework for refining and validating such information. Specifically, the framework will provide mechanisms for achieving the following tasks:

1. Annotation of Meta Information

   To refine the information from crowds, we have to collect meta information, such as information type, information source, and the time and location of corresponding events, associated with each crowd input. The information type refers to the intention of a message, e.g., situation update, call for help, ask for information about particular person, event, or location. The information source is also an key element because inputs from crowds may be mostly second-hand information that they received from other websites or even mass media. As we will have to pay more attention to first-hand information, a categorization of information source is required. Also, the time and location associated with those events must be explicitly annotated, because they may be different from the message posting time and the physical of poster.

2. Verification of Social Reports

   One main part of our plan is to estimate the trustworthiness and timeliness of crowd inputs. For messages without a reliable and traceable source and sufficient supports from other information, we plan to develop mechanisms which enable competent individuals to do information verification in a real-time fashion. The mechanisms will involve a message filtering scheme which picks most needed unverified messages from a pool according to the knowledge of a volunteer and a cheat-detection scheme that prevents misbehavior of volunteers. More concretely, the research issues at least include a) *information representation*, which decides how the crowd-contributed information (normally in a huge volume) is best appropriately presented to problem solvers; b) *task formulation*, which involves how to "knock" the knowledge out of volunteers in a most efficiently and error-free way; and c) *incentive provisioning*, which addresses the design to encourage active contributions to our tasks even if there will be no reward for the volunteers.

PERSPECTIVES ON FUTURE DIRECTIONS

Balance between Incentive and Quality

In my opinion, how to adequately provide sufficient incentives to workers while keeping reasonable outcome quality remains one of the most important issues in the human computation area [8]. I find a strong need to model the relationship between the provided incentive (e.g., monetary reward) and outcome quality for a human task. Doing so requires us

[1] http://teldap.tw/en/

to model the difficulty of a task beforehand, which is itself challenging as it is related to a human solver's ability, experience, and subjective perception. Nonetheless, balancing incentive and quality for human tasks is extremely critical to a healthy human computation environment and worths further investigation.

Design Pattern and Toolchain

To enable researchers and practitioners in various areas to solve their own AI-hard problems using a human computation approach, a well compiled set of design patterns [4], and even a chain of development toolkits, would be highly helpful. Ideally, there could a set of recipes, each of which can be facilitated by a wizard-like user interface, for newcomers to easily transform their problems into a sequence of human solvable tasks. I believe this research area would be key to the general use of human computation in fields.

Beyond Corwdsourcing Platforms

Currently, most crowdsourcing tasks are advertised and performed on dedicated crowdsourcing platforms such as Amazon Mechanical Turk and CrowdFlower. Therefore, the workers firstly need to have an account on such platforms and, secondly, they have to browse over the task lists repeatedly in order to find an appropriate task to solve. I am thinking the possibility to do *placement crowdsourcing* by placing human tasks on platforms which support small-amount currency exchange, such as online games and Apple iPad. This would involve the standardization of exchange formats, the dispatch of human tasks according to users' context and preferences, and implementation issues as multiple platforms/ devices/ software are involved. Nevertheless, efforts toward this direction could expand the size of candidate workers by orders of magnitude and eventually increase the potentials of the crowdsourcing approach in solving various problems.

AUTHOR'S BIOGRAPHY

Dr. Sheng-Wei Chen (also known as Kuan-Ta Chen) is an associate research fellow at the Institute of Information Science and the Research Center for Information Technology Innovation (joint appointment) of Academia Sinica. He received his Ph.D. in Electrical Engineering from National Taiwan University in 2006, and his B.S. and M.S. in Computer Science from National Tsing Hua University in 1998 and 2000, respectively. Prior to taking his academic path, he has been active as a programmer specialized in Windows and system programming, a technical writer of four books, a technical lecturer of programming courses, and a shareware developer.

His research interests include Internet and multimedia quality of experience (QoE) management, Internet measurement, network security, and online games. Much of his recent work focused on QoE-aware multimedia system design, AI-hard problem solving using human computation, and the combination of both directions. He received a Best Paper Award (with Ieng-Fat Lam and Ling-Jyh Chen) in IWSEC 2008 and K. T. Li Distinguished Young Scholar Award from ACM Taipei/Taiwan Chapter in 2009. He also received the Outstanding Young Electrical Engineer Award from The Chinese Institute of Electrical Engineering in 2010. He is a member of ACM, IEEE, IICM, and CCISA.

REFERENCES

1. Chen, K.-T., Chang, C.-J., Wu, C.-C., Chang, Y.-C., and Lei, C.-L. Quadrant of Euphoria: A crowdsourcing platform for QoE assessment. *IEEE Network* (2010).
2. Chen, K.-T., Wu, C.-C., Chang, Y.-C., and Lei, C.-L. A crowdsourceable QoE evaluation framework for multimedia content. In *Proceedings of ACM Multimedia 2009* (2009).
3. Chen, L.-J., Wang, B.-C., Chen, K.-T., King, I., and Lee, J. H.-M. An analytical study of puzzle selection strategies for the ESP game. In *IEEE/WIC/ACM Web Intelligence 2008* (2008).
4. Gamma, E., Helm, R., Johnson, R., and Vlissides, J. *Design patterns: elements of reusable object-oriented software*, vol. 206. Addison-wesley Reading, MA, 1995.
5. Ho, C.-J., Chang, T.-H., Lee, J.-C., Hsu, J. Y.-j., and Chen, K.-T. KissKissBan: a competitive human computation game for image annotation. In *HCOMP'09: Proceedings of the ACM SIGKDD Workshop on Human Computation* (2009), 11–14.
6. Ho, C.-J., and Chen, K.-T. On formal models for social verification. In *HCOMP'09: Proceedings of the ACM SIGKDD Workshop on Human Computation* (2009), 62–69.
7. Lin, C.-W., Chen, K.-T., Chen, L.-J., King, I., and Lee, J. H.-M. An analytical approach to optimizing the utility of ESP games. In *IEEE/WIC/ACM Web Intelligence 2008* (2008).
8. Mason, W., and Watts, D. Financial incentives and the performance of crowds. *ACM SIGKDD Explorations Newsletter 11*, 2 (2010), 100–108.
9. Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, ACM (2010), 851–860.
10. Starbird, K., Palen, L., Hughes, A., and Vieweg, S. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, ACM (2010), 241–250.
11. Vieweg, S., Hughes, A., Starbird, K., and Palen, L. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems*, ACM (2010), 1079–1088.
12. Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. recaptcha: Human-based character recognition via web security measures. *Science 321*, 5895 (2008), 1465.