Fast, Accurate, and Brilliant: Realizing the Potential of Crowdsourcing and Human Computation

Kurt Luther
School of Interactive Computing & GVU Center
Georgia Institute of Technology
Atlanta, GA, USA
luther@cc.gatech.edu

INTRODUCTION

Albert Einstein is attributed the famous inspirational quote, "Computers are incredibly fast, accurate, and stupid. Human beings are incredibly slow, inaccurate, and brilliant. Together they are powerful beyond imagination." Crowdsourcing and human computation may be among the most promising new approaches for combining the unique, often complementary, strengths of people and computers. Like many researchers, I am interested in exploring what those strengths are and how we might build systems to combine them in new, interesting, and useful ways. In this workshop paper, I describe several of my projects in this space and offer a vision for future research efforts.

RELATED PROJECTS

One way of thinking about my research is to return to the Einstein quote mentioned above. Two of my recent projects, one completed and one ongoing, take advantage of situations where human computation is *faster* and more *accurate* than the current state of purely computational options. Another pair of recent projects, one completed and one ongoing, use crowdsourcing to harness the *brilliance* that can emerge when people collaborate on creative projects. As a group, these four projects attempt to realize some of the potential for fast, accurate, and brilliant results hinted at by Einstein.

In Pursuit of "Fast" and "Accurate": Leveraging Human Computation for Making Complex Judgments

Two early human computation experiments, the ESP Game [1] and NASA's Clickworkers, identified *image recognition* as one human competency that, when aggregated via computational systems, can produce useful results. But image recognition is just one of many types of complex judgments humans can make, suggesting fertile ground for more experiments leveraging other human competencies. The following projects explore two additional competencies: transcribing speech (Audio Puzzler) and detecting improvisation in a performance (Digital Improv).

Audio Puzzler (completed)

The world, and especially the internet, is saturated with video and audio recordings of speech, most of which lack any transcription. Without transcription, however, it's hard to know what the content is about without listening to or watching all of it, or relying on coarse, manually-added metadata like tags and categories. The best automatic speech recognition (ASR) systems, while useful, produce error rates of 20-45% in real-world situations. Humans are experts at recognizing speech—can a human computation system produce better results at a reasonable cost?

Nicholas Diakopoulos, Irfan Essa, and I designed Audio Puzzler [7] to address this question. Although people can be quite good at transcribing speech, most are unmotivated to do so for very long. Hence, Audio Puzzler is designed as a casual Flash game. Just as the ESP Game motivated image tagging with a ludic (game-like) interface, so does Audio Puzzler motivate speech transcription with a ludic interface. However, the designs of the two games are quite different. As the name suggests, Audio Puzzler is a puzzle game and is not directly cooperative or competitive. The player is shown an interface with a series of bubbles; each bubble plays a chunk of audio (speech) when clicked. To complete each stage, the player must transcribe the speech for each bubble and assemble the bubbles in the proper sequence. The game awards points for speed and accuracy, and players can view a leaderboard of top scores.

Behind the scenes, Audio Puzzler aggregates the transcripts and also calculates timings for each word, which is useful for aligning the transcript with the audio/video source. Our evaluation of Audio Puzzler with 10 participants in a lab setting produced error rates of 10%, twice as good as the best ASR systems under normal conditions. Most participants offered positive feedback on the gameplay experience, though the type of speech content affected motivation and fun.

URL: http://www.audiopuzzler.com/

Digital Improv (in progress)

I've been working with Brian Magerko and several other researchers at Georgia Tech to investigate the cognitive processes of improvisational theater performers [11]. The goal of this research is to model these processes and contribute to making AI-based agents, e.g., robots or video game characters, behave in more dynamic, realistic ways.

After talking with professional improvisers and observing many performances, we noted that audiences react in very different ways depending on how improvised vs. scripted they believe a performance to be. This makes intuitive sense: the more time performers have time to plan (script) and practice a scene, the higher the expectations of the audience might be. However, we couldn't find empirical proof or quantification of this theory, so we sought to study it ourselves, using human computation.

In our study, conducted via Mechanical Turk, turkers were shown a short video of a performance and asked a series of questions about it. First, they were asked a test question about the content of the video. Then, they were asked to use a slider UI widget to rate the video along a series of dimensions, including how scripted, improvised, funny, dramatic, entertaining, believable, and easy to follow it was. After several pilots, we ran two batches of ≈150 HITs each.

In our data analysis—still in progress—we're examining the correlations between turkers' measures of improv-ness or scripted-ness, and other, quality-based metrics like humor and entertainment value. Beyond testing our theory mentioned above, we're also considering possible design implications. For example, online video hosts like YouTube may wish to categorize content by performance type (e.g., scripted TV show, improvised play) to provide context and set expectations for viewers.

In Pursuit of "Brilliant": Leveraging Crowdsourcing for Collaborative Creativity and Innovation

When listing the achievements of crowdsourced creativity, two examples stand out: Wikipedia and open source software, both developed by volunteers from around the world meeting and collaborating via the Internet. If these efforts can produce the world's largest encyclopedia and some of its best software, what else might crowdsourcing accomplish in other domains? To address this question, I turned to two domains filled with creative opportunities: scientific research (Pathfinder), and movie production (Pipeline).

Pathfinder (completed)

One of the earliest and most promising applications of crowdsourcing technology is in the domain of citizen science, the Clickworkers project mentioned above being just one example. The original conception of citizen science, beginning with the Christmas Bird Count in 1900, holds that ordinary citizens collect data for professional scientists to analyze, more cheaply and of a greater diversity and quantity than the scientists could collect themselves [6]. With the Pathfinder project, we wanted to go further, asking if citizens could also take part in the *analysis* of those data, i.e., crowdsourcing scientific knowledge production.

I worked with Scott Counts and his colleagues at Microsoft Research to develop Pathfinder [10], a web-based collaboration tool with two main feature sets. The first is the ability to upload, visualize, and share time-series data sets, called Tracks. The second is the ability to collaboratively analyze those Tracks with a feature called Discussions. Each Discussion begins with a Pathfinder user asking a question, e.g., "What is the relationship between traffic and pollution in Seattle?" Other users can join the Discussion and contribute in a variety of ways: adding background info, hypotheses, evidence, predictions, and to-dos. The collaborative analysis process is aided by two features. First, users can embed Tracks, including specific views and annotations, directly within the context of the Discussion. Second, the Discussion works like a structured wiki. Conversation can flow naturally, but key elements (e.g., evidence) can be tagged in a lightweight way and automatically summarized at the beginning of the Discussion. Any user can edit the Discussion, so errors can be corrected quickly and the organization can be improved over time.

For the "beta launch" of Pathfinder, we recruited 43 participants from commuting-oriented email lists and asked them to contribute data and discuss questions related to commuting and local transportation issues. We followed this with a more controlled user study (15 participants) comparing Pathfinder to a standard wiki. We found that participants preferred Pathfinder and were able to engage in deeper scientific analysis. We also saw a need to provide additional technological support, such as mechanisms for attribution, because participants were generating original research, not just summarizing existing knowledge.

URL (the Discussion features mentioned above are not yet public): http://datadepot.msresearch.us/

Pipeline (in progress)

As tools for movie and game production grow cheaper, more powerful, and easier to use, models of production are evolving to accommodate an influx of passionate novices and amateurs. For example, the Mass Animation project paired amateur animators recruited via Facebook with a team of professional filmmakers to create "Live Music" (2009), a 3D animated short film screened in theaters across the United States [2]. This hybrid model was highly successful, begging the question: is it possible to crowdsource every aspect of a similar movie production?

Amy Bruckman and I conducted an initial study which found that people were already collaborating over the Internet on animated movie projects called "collabs" [8]. Collabs marry crowdsourcing with elements of traditional creative collaboration. To start, a leader proposes a movie idea, usually one that can be modularized into many independent pieces and recomposed later, e.g., a story with 10 chapters. Artists claim a chapter, animate it, and submit it to the leader, who assembles each piece into a single, coherent movie. Unlike open source projects, artists claim ownership and authorship over their work and leaders must carefully negotiate change requests.

Although many finished collabs demonstrate an impressive level of creativity and craftsmanship, most collabs are never completed, contrary to the wishes of their members [9]. We have built Pipeline, a web-based system for supporting and enabling new types of successful collabs. Our early work found that many obstacles to collab success center around overburdened leaders, so Pipeline's focus is easing the burden on leaders, through automation, decentralization, and improved group awareness. A key innovation in Pipeline is the notion of "trust": creators of new collabs can choose to trust only themselves or a small group of leaders with Pipeline's advanced features, or trust all project members by default. This latter, more wiki-like approach is aided by a complete history of user actions; low quality contributions can be undone at any time with a single click.

Pipeline is launching in early 2011 and will eventually be made open source. We are running a series of contests where half of participants are assigned to use Pipeline for a collab, and half use traditional methods. By analyzing logs and interviewing Pipeline users at the conclusion of each contest, we hope to learn about how different leadership styles and technological supports affect the process and outcome of crowdsourced creativity.

ENVISIONING THE FUTURE

I'm excited about the future of human computation and crowdsourcing, which seems very promising. Although this research field is a young one, I've already observed a wide variety of inspiring and successful projects at CHI, CSCW, and other venues. Drawing on my familiarity with these projects and the work described above, I offer the three ideas as potential focus areas for future research.

New Domains

Humans and computers can be good at many things, and the ways the two can be combined are almost limitless. I envision a future in which the domains of crowdsourcing and human computation are almost as vast and varied as those of human endeavors generally. Almost certainly, these techniques will not be appropriate for every task in every domain. They are, however, likely to be useful for a few tasks in most domains. One of the key challenges in contributing to new domains is adapting to problems that are more open-ended and less well-defined, especially in design-oriented fields [5]. However, we must face these challenges before we can deal with them, and solutions in one domain may transfer to others. The wider researchers cast their experimental nets, the clearer our collective picture will be of the breadth and limitations of crowdsourcing and human computation.

New Complexities

As more crowdsourcing and human computation experiments show success in aggregating many simple tasks, an important next step is to increase the complexity of the tasks these systems set out to accomplish. Already, researchers have developed systems that label images [1], transcribe speech [7], proofread papers [4], and provide many other valuable services. But what about systems that draw images, give speeches, and write papers? The Find-Fix-Verify pattern [4] offers a promising framework for workers to build upon the efforts of other workers, a crucial step towards achieving more complex goals. However, much more can still be done. I envision a future in which crowdsourcing becomes a primary means by which content is created, not just edited, annotated, or identified. Benkler observes that peer production is most successful when people can self-select tasks that are not only modular, but also hetereogeneously granular, allowing them to match the task to their available time, effort, and interest [3]. The best crowdsourcing systems will support workers with diverse motivations, skills and abilities, and offer a range of tasks, from simple to complex.

New Literatures

Years ago, when online collaboration and virtual teams emerged as viable ways to conduct business, they challenged HCI/CSCW researchers, who sought out theory and research from other disciplines to help explain what was going on. They also challenged scholars in those disciplines, who had to refine or expand some of their own theories to account for new ways of working. Similarly, human computation and crowdsourcing bring a new set of challenges that once again cause us to question our assumptions about motivation, leadership, fairness, and the technologies that foster or hinder them. As researchers, we must be willing to consider a wide range of perspectives beyond the canonical HCI/CSCW literature. Multi-disciplinary collaborations are essential to producing strong results that are applicable beyond our own fields.

CONCLUSION

Crowdsourcing and human computation can produce results that are fast, accurate, and brilliant. Many current projects that I've seen already do. However, I believe that we have only seen the tip of the iceberg in terms of realizing the potential of these techniques. We must expand research to new domains beyond what has already been tried. We must pursue tasks of increasing complexity and design systems that are sensitive to the diverse motivations, skills, and abilities of the crowd. And finally, we must revisit our own assumptions about HCI, CSCW, and online collaboration, and strive to integrate new perspectives from other literatures that deepen our own understanding. The future for crowdsourcing and human computation is bright, as long as we continue to raise our expectations and, just as importantly, stretch our imaginations.

AUTHOR BIOGRAPHY

Kurt Luther is a Ph.D. candidate in Human-Centered Computing at the Georgia Institute of Technology. His advisor is Dr. Amy Bruckman. Kurt designs and studies social computing systems that support creative collaboration. For his dissertation work, he is building and launching a Pipeline, a platform for crowdsourcing movies, to examine the role of leaders in online creative collaboration. He received his bachelor's degree in 2006 from Purdue University, where he studied computer graphics, art, and design. He has also interned at YouTube, Newgrounds, Microsoft Research, and IBM Research.

ACKNOWLEDGMENTS

Thanks to my study participants and my collaborators on the aforementioned research projects, especially Amy Bruckman, Scott Counts, Nicholas Diakopoulos, and Brian Magerko. This research is supported by National Science Foundation grant CreativeIT-0855952.

REFERENCES

1. von Ahn, L. and Dabbish, L. Labeling images with a computer game. *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM (2004), 319-326.
2. Barnes, B. An Animated Film Is Created Through Internet Consensus. *The New York Times*, 2009. http://www.nytimes.com/2009/07/16/movies/16mass.html?_r=1.
3. Benkler, Y. Coase's Penguin, or Linux and the Nature of the Firm. *Yale Law Journal 112*, 3 (2002), 369-446.
4. Bernstein, M.S., Little, G., Miller, R.C., et al. Soylent: a word processor with a crowd inside. *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, ACM (2010), 313–322.
5. Buchanan, R. Wicked Problems in Design Thinking. *Design Issues 8*, 2 (1992), 5-21.
6. Cohn, J.P. Citizen science: can volunteers do real research? *BioScience 58*, 3 (2008), 192-197.
7. Diakopoulos, N., Luther, K., and Essa, I. Audio Puzzler: piecing together time-stamped speech transcripts with a puzzle game. *Proceeding of the 16th ACM international conference on Multimedia*, ACM (2008), 865-868.
8. Luther, K. and Bruckman, A. Leadership in online creative collaboration. *Proceedings of the ACM 2008 conference on Computer supported cooperative work*, ACM (2008), 343-352.
9. Luther, K., Caine, K., Ziegler, K., and Bruckman, A. Why it works (when it works): success factors in online creative collaboration. *Proceedings of the 16th ACM international conference on Supporting group work*, ACM (2010), 1–10.
10. Luther, K., Counts, S., Stecher, K.B., Hoff, A., and Johns, P. Pathfinder: an online collaboration environment for citizen scientists. *Proceedings of the 27th international conference on Human factors in computing systems*, ACM (2009), 239-248.
11. Magerko, B., Manzoul, W., Riedl, M., et al. An empirical study of cognition and theatrical improvisation. *Proceeding of the seventh ACM conference on Creativity and cognition*, ACM (2009), 117–126.