Personalized Online Education — A Crowdsourcing Challenge

Daniel S. Weld[1] **Eytan Adar**[2] **Lydia Chilton**[1] **Raphael Hoffmann**[1] **Eric Horvitz**[3]
Mitchell Koch[1] **James Landay**[1] **Christopher H. Lin**[1] **Mausam**[1]

[1] Department of Computer Science & Engineering, University of Washington, Seattle, WA 98195
[2] School of Information, University of Michigan, Ann Arbor, MI 48109
[3] Microsoft Research, Redmond, WA 98052

Abstract

Interest in online education is surging, as dramatized by the success of Khan Academy and recent Stanford online courses, but the technology for online education is in its infancy. Crowdsourcing mechanisms will likely be essential in order to reach the full potential of this medium. This paper sketches some of the challenges and directions we hope HCOMP researchers will address.

Introduction

Educators, parents, and students have long dreamed of the possibility of personalized education — teaching individual students according to their abilities, interests, motivations, and learning styles. To date, such personalized education has been limited by scarce human resources, but perhaps technology offers a cure. The idea of large-scale, online courses conjures images of impersonal, mediocre, and standardized coursework for the masses, but in contrast we believe that there are vast opportunities for engaging "the crowd" to personalize education and scale high-quality tutoring via online engagement. Indeed, we argue this is a grand challenge for researchers in human computation.

By itself, crowdsourcing is unlikely to deliver the best educational experience, but it is a natural framework for learning. Online tutoring systems have made considerable progress in recent years, and one can envision them as "computational" agents in a mixed human-computer social system. Machine learning and datamining form another crucial component; by analyzing patterns from the behavior of a myriad of students one may learn latent variables that predict the efficacy of different instructional techniques for specific individuals. Similar techniques may connect struggling students with appropriate tutors and recommend which questions deserve immediate attention.

Planning methods, which have been shown to improve the efficiency of workflows on labor marketplaces such as Mechanical Turk, may prove useful for optimizing personalized curricula (Dai, Mausam, and Weld 2010; Shahaf and Horvitz 2010; Dai, Mausam, and Weld 2011; Lin, Mausam, and Weld 2012). Instead of minimizing the cost of accomplishing a content creation or translation task, the objective might be to maximize the expected student competence across a range of topics while holding constant the time spent on videos and worked-problems.

Developing crowdsourcing methods for improved education will require numerous technical advances from many researchers. In the next section, we argue that now is the time for a concerted push in this area, and highlight some areas of recent progress. We conclude by highlighting a set of challenges for future work.

Why Crowdsourcing & Education?

We see three reasons why education is an exciting direction for crowdsourcing research: 1) crowd-techniques will be required in order to deliver quality education in some areas; 2) existing techniques are ready for application to this new area; and 3) online education represents a new, relatively-unexplored way of creating crowds.

Scaling Creative Education Requires a Crowd

Most existing online education systems are restricted to multiple-choice or known-answer questions, because they are easy to grade. Programming assignments can be auto-graded, but only when the requirements are very narrowly defined. The Educational Testing Service (ETS) uses natural language processing (NLP) algorithms to augment human assessments of English fluency for essay questions (Monaghan and Bridgeman 2012), but such automated methods require too much investment for most course situations. So how can one teach creative processes (e.g., English composition or user interface design) where criteria change from year to year, evaluation is subjective and feedback is essential? One could use a marketplace like oDesk [1] to hire teaching assistants, but even so, one would need to use crowdsourcing techniques to standardize feedback across thousands of TAs and monitor their performance.

An obvious alternative might be to ask students to evaluate each other's work. It seems intuitively plausible that one could use the crowd to evaluate open-ended questions with high reliability. Indeed, previous studies suggest that,

[1] odesk.com

when guided by a clear rubric, peer-assessment has extremely high correlation with teacher-assigned grades (e.g., $r > 0.91$) (Sadler and Good 2006). Crowdsourced peer-grading may also lead to more accurate assessments of current and future performance by combining the opinions of graders with diverse perspectives, expertise, and stakes in a students progress (Page 2008). Furthermore, self-grading was found to increase student learning; unfortunately, however, results are mixed concerning the benefit of peer-grading on the peer reviewers (Sadler and Good 2006; Cho and Cho 2011).

Automating the peer-evaluation process involves several challenges, which we outline below, but the potential benefit is high. By finding students with different perspectives to evaluate a single work, we may be able to achieve an even higher-quality evaluation compared to a single instructor (who typically has limited time and a single perspective).

Crowdsourcing Methods can Improve Education

Several techniques, pioneered in the computer-supported cooperative work (CSCW) and crowdsourcing communities offer potential benefits to online education, and some methods have already been directly applied.

Many researchers have proposed methods for determining the accuracy of workers in online labor markets, such as Mechanical Turk (Whitehill et al. 2009; Welinder et al. 2010; Dai, Mausam, and Weld 2011; Lin, Mausam, and Weld 2012; Wauthier and Jordan 2012). *Item response theory* (Hambleton, Swaminathan, and Rogers 1991) uses similar strategies to measure the reliability of standardized exams and to normalize scores between successive versions of an exam. We will wish to extend these algorithms for online education to measure student competence on multiple dimensions. Accurate measures of student competence will inform not just which problems and content materials to suggest for a student, but might also recommend likely tutors from the pool of more advanced students. In order to support rapid question-answering, real-time crowdsourcing techniques may prove popular (Bernstein et al. 2011).

A flexible analytic platform will underly student tracking, confusion detection, curriculum optimization, question routing and other functions. Clustering techniques, originally developed for understanding the behavior of website visitors, also apply to navigation patterns across courseware (Cadez et al. 2000). Andersen *et al.* (2010) introduce such a method for clustering and visually summarizing student traces in educational games, independent of the game's structure. Coursera[2] has analyzed student traces to determine which online videos are rewatched and in what order as well as to determine which forum posts, when read, are likely to lead to improved homework submission (Koller 2012).

Statistical A/B testing, long used to optimize user interactions on e-commerce websites (Kohavi, Henne, and Sommerfield 2007), could be applied to determine the most effective way to present online course materials. Some initial efforts in this direction have already been deployed. For example, Andersen *et al.* (2011) varied the motivation scheme in two online educational games, using A/B testing across 27,000 players to determine the effect on the time students spent playing the game. They observe that secondary objectives, which conventional wisdom deemed to increase a game's replayability and depth, can actually cause many players to play for less time — unless the secondary objectives are carefully designed to reinforce the game's primary goal. Work on adaptive websites and self-personalizing interfaces should inform the design of self-adjusting curricula (Perkowitz and Etzioni 2000; Anderson, Domingos, and Weld 2002; Gajos et al. 2006).

Socially-organized crowds are also important for education, but not only for their obvious potential to increase motivation and commitment in classes. Consider immersion language learning, which can be extremely frustrating if one is seldom understood, and doesn't understand what errors are being made. Electronic dictionaries may help a learner cope with real-life situations, but help little with learning. Translater, a mobile phone application, records one's troubled attempts to speak Chinese in public, uploads an audio recording to a server, and allows friends (or paid tutors) to diagnose one's interactions and provide instruction (Chilton and Landay 2012).

Online Education Can Draw a Crowd

Traditionally, there have been three main ways of assembling a crowd to accomplish a task: pay them, e.g., Mechanical Turk, entertain them, e.g., the ESP Game (von Ahn and Dabbish 2004) and FoldIt (Cooper et al. 2010), or create a community, e.g., Wikipedia and Stack Overflow. Now it is clear that by offering an online course one can attract a crowd of hundreds of thousands or more students. And, uniquely, this is a skilled crowd, one which has at least the basic skill sets and prerequisites for the course. Can their joint activities be leveraged to improve educational materials and the process of education itself?

Duolingo is one success story; by attracting students to practice a foreign language, they improve language resources with new translations. In-place collaborative document-annotation systems offer a different approach (Brush et al. 2002; Zyto et al. 2012) — encouraging students to discuss and elaborate course content in the context of an online textbook. In order to reduce fragmentation, Coursera employs a scheme where students, before they can submit their question, are prompted with previous posts that may address a similar topic (Koller 2012). These comments, like replaying parts of online videos and "like" votes for forum answers, promise to increase the clarity of the curriculum over time.

A different idea would be to use machine learned classifiers and information extraction methods to find educational materials (syllabus pages, problem sets, videos, definitions, tutorials, slides, etc.) on the Web and integrate them on an education portal site. If the underlying Web crawling and extraction methods were good enough to create a useful resource, then the site would generate native traffic (a crowd) which could be encouraged to provide comments to improve

[2] www.coursera.com

the site, creating positive feedback (Hoffmann et al. 2009). Reward mechanisms (such as those used by Stack Overflow) could also be helpful.

Challenges

It is clear that crowdsourcing methods have great potential for improving online personalized education. The following challenge areas offer the promise of rapid progress and substantial impact.

Content Creation & Curation

To date, curriculum design for online courses (whether Khan Academy, Coursera or Udacity) has been centralized, but the success of Wikipedia shows that community action can create incredible resources. Indeed, there are already a large number of educational resources on the Web, including slides, problem sets and video lectures. How can the process of creating educational content be assisted by the crowd?

Some of these problems have already been partially addressed — for example, by the development of Web-based, open authoring tools (Aleahmad, Aleven, and Kraut 2009), but extending these systems to support thousands (or more) of contributors of varying competence will require considerable work. Discussion forums are an obvious first target. Stack Overflow, reddit and similar sites have developed rating systems that draw attention to good questions and answers. But we suspect that all these systems will bog down with a morass of disjoint questions and answers over time. How can similar questions get distilled and summarized? How can source material be revised in order to increase clarity and obviate the need for individual questions. Perhaps mechanisms such as those introduced by the political discourse system Considerit (Kriplean et al. 2012), could be adapted to the education context. Another possibility might be to develop a recommender system, like Wikipedia's SuggestBot (Cosley et al. 2007), that could route editing tasks to appropriate volunteers. This approach might work well for transcribing lectures and videos into different languages, adding links to research papers, or rewriting text to make terminology consistent or language easier to understand.

Other types of course content are more obviously decomposable and should be easier to crowdsource: alternate examples, improved figures, polished animations, and additional problem sets. The traditional crowdsourcing problem of quality control will apply to the creation of most new content. Iterative improvement (Little et al. 2009; Dai, Mausam, and Weld 2010) and find-fix-verify workflows (Bernstein et al. 2010) should be useful here. One might even imagine applying Soylent-like techniques (Bernstein et al. 2011) to allow a professor to get immediate feedback on an explanation or example from a real-time test crowd, before delivering the polished presentation to the full class.

Personalization & Engagement

Today, most online courses are following what brick-and-mortar schools have practiced for ages: one-to-many presentations with poor interactive learning content. In contrast, the online medium offers potential for the exact opposite: adaptive one-to-one interaction with the opportunity to directly engage students.

One can imagine personalizing many aspects of a course: textbook reading, video segments, forum posts, challenge problems and discussion partners (dynamic tutors). Underlying any such effort will be machine learning algorithms for tracking a students skills and abilities, gauging their motivation to study, and predicting when a student is stuck on a concept and needs help. In some subjects, e.g., early mathematics and physics, the space of possible student misconceptions is well articulated; here, simple classification-based approaches may be applicable. In other subjects, more sophisticated, unsupervised learning methods will be necessary. Student modeling has been extensively studied in the literature on intelligent tutoring systems (Woolf 2009), but online education completely changes the magnitude of available training data.

Improving long-term engagement is an important area for future work. Less than 15% of students completed the Norvig/Thrun online AI class; only 8% made it through MIT's recent class (Lewin 2012) and of the 104,000 students who signed up for Stanford's 2011 Machine Learning class, 46,000 submitted the first assignment, and 13,000 received a passing grade (Koller 2012). Integrating online courses with a student's social network may help. Some have proposed online hangout environments and methods for linking students with common interests yet disparate skill sets, but how can one build and scale such systems? Dynamic tutor pairings may benefit the tutor as much as the tutee, but will students embrace such artificial connections? Reputation systems that reward students who tutor others could enable both a sense of pride among tutors and an overall improved experience of education for all.

There is a fundamental tension between social reinforcement and the potential benefit of anytime-delivery of education. Should online courses be synchronous or asynchronous? Coursera and Udacity are embracing the synchronous model in which a cohort of students progress through the curriculum in lockstep (Koller 2012). There appear to be clear benefits to this approach — shared deadlines may improve motivation, and a critical mass on forums improves the chances of quick answers. Stanford's 2011 Machine Learning class had an impressive 22 minute median response time for forum questions (Friedman 2012). On the other hand, an asynchronous delivery model offers so much additional freedom for the learner that the alternative feels contrived. Can one develop a social mechanism that combines the best of both worlds? How would one introduce group activities if everyone is working at different times? Could students "engage" with others who took the class previously (e.g., by introducing ESP-game-like delays, which make one think one is interacting with someone live when one is basically playing against a recorded session (von Ahn and Dabbish 2004))?

In any case, social mechanisms are likely only part of the story. We need to develop multiple ways to engage and hook different types of learners. Rating and reputation systems ("badges") may work for some students, but they are likely only a smal part of the solution. Online courses can

be considered laboratories for studying incentive programs (including games) that enhance learning and teaching.

Providing Rich Feedback to Students

Humans are likely necessary for providing feedback on subjective, creative processes (e.g., English composition, user interface design or software architecture for an open-ended project). Can one design crowdsourced workflows that produce quality feedback for this sort of coursework from workers (students) whose understanding of the material is imperfect? How much redundancy is necessary to achieve consistency if grading were performed this way? Finally, can peer assessment be designed in a way that confers educational benefits to the participants, rather than merely using them as a source of free labor?

We conjecture that today's understanding of workflow design (e.g., iterative improvement and find-fix-verify) will suffice for providing useful feedback for small assignments, e.g., essays of a page or less, but are inadequate for assessing larger blocks of work in which an integrated perspective is essential.

Reward systems constitute another challenge — how can one induce students to provide feedback to their classmates? On first blush, it appears that open voting schemes, such as those used on sites like Stack Overflow might be adapted. But there is a key difference — those sites thrive when only a small percentage of visitors actively participate. We conjecture that such imbalance is unsustainable in the education context in which every student needs detailed feedback.

Simply requiring students to evaluate their peers as part of their grades also has limitations. Anecdotal evidence suggests that students may act irrationally in the face of these incentives. Furthermore, is such a scheme fair as part of assessment? Providing feedback requires different skills than simply learning material to the point of mastery. Finally, it is harder to provide constructive feedback for work with few flaws than for work with obvious deficiencies. Here again, learning algorithms may offer benefits by tracking students abilities to answer questions, modeling their skill at grading the answers of others, and routing grading jobs to the right student.

Conclusion

Personalized online education is an exciting domain rife with challenges for crowdsourcing research. We have the power to transcend the one-to-many broadcast model of education for one-to-one adaptive methods and personalized interactive learning. To realize this potential we must develop scalable assessment models that transcend multiple choice, testing creativity and recall, not just recognition; crowdsourced methods, such as co-grading and student curation of curricular content will be necessary. Multi-dimensional student modeling may drive these innovations, and it seems a good approach might extend the worker-accuracy models originally developed for crowdsourced labor markets. We foresee exciting developments in crowdsourced methods for content creation & curation, personalization and generation of rich feedback for students.

Acknowledgements

We thank Alan Borning, Jonathan Bragg, Daphne Koller, Zoran Popovic, and Noah Smith for insightful discussions. This work was supported by the WRF / TJ Cable Professorship, Office of Naval Research grant N00014-12-1-0211, and National Science Foundation grant IIS 1016713.

References

Aleahmad, T.; Aleven, V.; and Kraut, R. 2009. Creating a corpus of targeted learning resources with a web-based open authoring tool. *Learning Technologies, IEEE Transactions on* 2(1):3 –9.

Andersen, E.; Liu, Y.-E.; Apter, E.; Boucher-Genesse, F.; and Popović, Z. 2010. Gameplay analysis through state projection. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, FDG '10, 1–8. New York, NY, USA: ACM.

Andersen, E.; Liu, Y.-E.; Snider, R.; Szeto, R.; Cooper, S.; and Popović, Z. 2011. On the harmfulness of secondary game objectives. In *Proceedings of the 6th International Conference on Foundations of Digital Games*, FDG '11, 30–37. New York, NY, USA: ACM.

Anderson, C. R.; Domingos, P.; and Weld, D. S. 2002. Relational Markov models and their application to adaptive web navigation. In *Proc. of the 2002 Conference on Knowledge Discovery and Data Mining*.

Bernstein, M.; Little, G.; Miller, R.; Hartmann, B.; Ackerman, M.; Karger, D.; Crowell, D.; and Panovich, K. 2010. Soylent: A word processor with a crowd inside. In *UIST*.

Bernstein, M.; Brandt, J.; Miller, R.; and Karger, D. 2011. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *UIST*.

Brush, A. J. B.; Bargeron, D.; Grudin, J.; Borning, A.; and Gupta, A. 2002. Supporting interaction outside of class: anchored discussions vs. discussion boards. In *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community*, CSCL '02, 425–434. International Society of the Learning Sciences.

Cadez, I.; Heckerman, D.; Meek, C.; Smyth, P.; and White, S. 2000. Visualization of navigation patterns on a web site using model-based clustering. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 280–284. New York, NY, USA: ACM.

Chilton, L., and Landay, J. 2012. Personal communication.

Cho, Y. H., and Cho, K. 2011. Peer reviewers learn from giving comments. *Instructional Science* 39(5):629–643.

Christian, C.; Lintott, C.; Smith, A.; Fortson, L.; and Bamford, S. 2012. Citizen science: Contributions to astronomy research. *CoRR* abs/1202.2577.

Cooper, S.; Khatib, F.; Treuille, A.; Barbero, J.; Lee, J.; Beenen, M.; Leaver-Fay, A.; Baker, D.; Popovic, Z.; and players, F. 2010. Predicting protein structures with a multiplayer online game. *Nature* 756–760.

Cosley, D.; Frankowski, D.; Terveen, L.; and Riedl, J. 2007. Suggestbot: Using intelligent task routing to help people find

work in wikipedia. In *Proceedings of the 2007 Conference on Intelligent User Interfaces*.

Dai, P.; Mausam; and Weld, D. S. 2010. Decision-theoretic control of crowd-sourced workflows. In *AAAI10*.

Dai, P.; Mausam; and Weld, D. S. 2011. Artificial intelligence for artificial, artificial intelligence. In *AAAI*.

Friedman, T. 2012. Come the revolution. `http://www.nytimes.com/2012/05/16/opinion/friedman-come-the-revolution.%html`.

Gajos, K. Z.; Czerwinski, M.; Tan, D. S.; and Weld, D. S. 2006. Exploring the design space for adaptive graphical user interfaces. In *AVI '06: Proceedings of the working conference on Advanced visual interfaces*, 201–208. New York, NY, USA: ACM Press.

Hambleton, R.; Swaminathan, H.; and Rogers, H. 1991. *Fundamentals of Item response Theory*. Sage Press.

Hoffmann, R.; Amershi, S.; Patel, K.; Wu, F.; Fogarty, J.; and Weld, D. S. 2009. Amplifying community content creation with mixed-initiative information extraction. In *Proceedings of the 2009 Conference on Human Factors in Computing Systems (CHI-09)*.

Kohavi, R.; Henne, R. M.; and Sommerfield, D. 2007. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *KDD*.

Koller, D. 2012. Personal communication.

Kriplean, T.; Morgan, J.; Freelon, D.; Borning, A.; and Bennett, L. 2012. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, 265–274. New York, NY, USA: ACM.

Law, E., and von Ahn, L. 2011. *Human Computation*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Lewin, T. 2012. `http://www.nytimes.com/2012/05/03/education/harvard-and-mit-team-up-to-%offer-free-online-courses.html`.

Lin, C.; Mausam; and Weld, D. S. 2012. Dynamically switching between synergistic workflows for crowdsourcing. In *AAAI*.

Little, G.; Chilton, L. B.; Goldman, M.; and Miller, R. C. 2009. TurKit: Tools for Iterative Tasks on Mechanical Turk. In *Human Computation Workshop (HComp2009)*.

Monaghan, W., and Bridgeman, B. 2012. E-rater as a quality control on human scores. `http://www.ets.org/Media/Research/pdf/RD_Connections2.pdf`.

Page, S. 2008. *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton University Press.

Perkowitz, M., and Etzioni, O. 2000. Towards adaptive web sites: Conceptual framework and case study. *Artificial Intelligence* 118:245–276.

Sadler, P., and Good, E. 2006. The impact of self- and peer-grading on student learning. *Educational Assessment* 1–31.

Shahaf, D., and Horvitz, E. 2010. Generlized markets for human and machine computation. In *Proceedings of the Twenty-Forth Conference on Artificial Intelligence*.

von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 319–326. New York, NY, USA: ACM.

Wauthier, F., and Jordan, M. 2012. Bayesian bias mitigation for crowdsourcing. In *In Proc. of NIPS*.

Welinder, P.; Branson, S.; Belongie, S.; and Perona, P. 2010. The multidimensional wisdom of crowds. In *NIPS*.

Whitehill, J.; Ruvolo, P.; Wu, T.; Bergsma, J.; and Movellan, J. 2009. Whose vote should count more: Optimal integration of labels from laberlers of unknown expertise. In *In Proc. of NIPS*, 2035–2043.

Woolf, B. P. 2009. *Building Intelligent Interactive Tutors*. Morgan Kaufmann.

Zyto, S.; Karger, D.; Ackerman, M.; and Mahajan, S. 2012. Successful classroom deployment of a social document annotation system. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, 1883–1892. New York, NY, USA: ACM.