Crowdsourcing the Research Process

Rajan Vaish
Computer Science Department
University of California at Santa Cruz
1156 High Street, Santa Cruz, California 95064, USA
rvaish@cs.ucsc.edu

Abstract

Research is a high skill and resource intensive activity, both in time and effort, and often follows an ad hoc process. In a research process, its often unclear what ingredients; or what recipe or process, which if repeated produces a publishable paper. Meanwhile, experienced researchers with novel ideas are constrained with limited time and funding resources; and motivated students with exceptional skill-sets lack direction or research mentor. In this proposal, I introduce a research direction which explores the possibility of expert crowdsourcing the research process, by connecting mentor with student crowd. The process would allow mentors to systematically use operators such as split, merge, remove or add on project ideas, code or students to manage research process and crowd. The process would include series of research phases like, brainstorming, paper-pencil prototyping, development and user-evaluation to produce publishable results. Encouraged by prior pilot experiment findings, my doctoral research examines the possibility of crowdsourcing the research process using operators along the research phase, while solving resource and opportunity constraints among mentor and crowd.

Keywords expert crowdsourcing, research methodology, education, collaboration, distributed research

Motivation

Conducting or continuing with novel research is a valuable activity; irrespective of the domain, it consistently furthers the state of knowledge possessed by humanity. However, research is a resource intensive activity requiring expert knowledge. For any particular type of research, its often unclear what reproducible processes can lead to a publishable paper. This is truer for inexperienced researchers, but even experienced researchers might follow an ad hoc protocol.

Researchers with a consistent publishing history at top organizations and universities often have many great ideas which they want to pursue further, but lack sufficient time and resources to execute them. Due to limited time, budget and funding, a researcher/professor can hire only a certain number of students/interns. At the same time, there are a large number of undergraduate and graduate students, all around the world, who want to get involved in a research project, but are not able to do so due to 1) Ignorance about the research process itself; 2) Unable to seek a mentor in their location; 3) Are good at certain skills (like programming, or running user studies), but dont know how to write a research paper.

What if there was a recipe to do research and produce publishable papers? What if busy researchers can simply crowdsource their research, and publish faster; therefore contributing to the world, and help in students career building worldwide? What if it was possible to mobilize student crowd and get them involved in lieu of a stellar CV? Would it possible to develop an algorithmic process to manage student crowd and research process? These questions motivated me to pursue this project as part of my PhD program.

Crowdsourcing enables individuals to come together and complete projects that would be virtually impossible for a single individual to accomplish at the same scale (Bernstein et al. 2010) (Kittur et al. 2011). To harness the potential of crowdsourcing, researchers and practitioners have often relied and successfully used crowdsourcing platforms or marketplaces such as Amazon Mechanical Turk. However, in spite of a large worker base and past successes, these platforms come with their own limitations. Among many constraints, it is difficult to accomplish tasks which require a broader participation base, or expert knowledge. Therefore, causing many creative, open-ended or highly complex tasks or research questions remain largely unsolved.

Motivated by these research questions, my research focuses on exploring the possibility of expert crowdsourcing by letting researchers manage student crowd and research process to accomplish the goal. The process would allow mentors to systematically use operators such as split, merge, remove or add on project ideas, code or students; along a series of research phases as part of the algorithmic process.

Background and Related Work

To accomplish tasks which require a broader participation base, or expert knowledge; researchers are tending away from crowdsourcing marketplaces, and moving towards the development of expert crowdsourcing projects. Expert crowdsourcing projects such as Ensemble (Kim, Cheng, and Bernstein 2014) and LeadGenius (leadgenius

2014) are already enabling us to achieve complex tasks which werent possible before.

Crowd can be creative (Yu, Kittur, and Kraut 2014) and has been put to use to foster prototyping process (Dow et al. 2010) and research process (Cranshaw and Kittur 2011). Research is a complex task, traditionally conducted in a small group. However, with the surge of crowdsourcing potential, researchers have attempted to author academic paper in a large distributed group (Tomlinson et al. 2012). Inspired by the success of MOOCs, researchers at UC San Diego aim to explore the area of MOOR (massive open online research) (MOOR 2013).

To foster the state of academic research and collaboration, Miller et. al. (Miller et al. 2014) at MIT and Northwestern University attempted to explore the concept of peer research. However, one wonders if it is possible to create an algorithmic approach to research process or paper writing. So far Jaime Teevan from Microsoft Research seems to be exploring the area (Formula 2013), but theres no published result to claim the success. I believe the research direction being examined in this proposal is novel and impactful.

Proposed Research

The proposed research about expert crowdsourcing of the research process attempts to address two major questions, and eventually aims to produce publishable paper or results as part of the experiment. Please note; by experiment I mean, experiment to test the efficacy of the research process implemented. To describe the proposed research, I would divide this section into two primary questions and define methodology as an investigational response to them:

Whether it is possible to connect researchers with student crowd, and incur motivation to solve resource and opportunity constraints among them? Experienced researchers have novel ideas, but fall short of time and funding resources. Self-motivated students serious about their career and interest are often highly skilled, but lack research direction or mentorship. To connect them, I plan to run a call for project at universities, where students can apply voluntarily (as an individual or as a team), and share information about their skills and time commitment. Students will have to or clubbed into a team to participate further. The projects research problem would be chosen by an experienced researcher, who would be willing to participate in the experiment, and commit upto one to two hour of their time per week. I believe that it is important to minimize researchers time commitment, while preventing it to affect projects progress - failure to do so might defeat the purpose.

Depending on the nature of the project, the process might take up to few weeks. Considering this fact, a proportionate number of students might drop off from the experiment. Therefore, making it important to consider recruiting extra students. Students who stay until the completion of the project will be added as co-authors of the paper to be produced of their work. The percentage of drop-outs and interviewing/surveying students from phase to phase can help us understand the motivational aspects surrounding this experiment, and possibly improve it based on the feedback.

Whether it is possible to develop an algorithmic process or recipe to manage research process and student crowd to produce publishable results or paper? The project proposed as part of this experiment would consist of a series of research phases such as: brainstorming ideas, paper-pencil prototyping, development, running experiments, generating results or conducting user-studies, analysis and paper writing. In order to have a streamlined progress across these phases, and for mentor to manage student crowd - I believe that an algorithmic process or recipe would help towards the success of the project. Inspired by the MapReduce algorithm (Dean and Ghemawat 2008) used on big data, I propose to use a similar algorithm to manage the student crowd and their progress to maximize productivity for students and researchers. The structure of the algorithm is stated below:

- **Merge**
  - What
    - Ideas/approaches
    - Student crowd/teams
    - Project code
  - When
    - Time constraints/fasten the process
    - Value addition or unite resources for productivity when similar idea for both teams.
    - To achieve bigger goal
- **Split** (to parallelize)
  - What
    - Ideas/approaches
    - Student crowd/teams
  - When
    - Want diversity in idea or approach or algorithm
    - To filter out unproductive resources from productive ones
    - Team creation process
    - Initial multiple idea generation process
    - A team too large to be productive
- **Add** (external)
  - What
    - External idea/alternative approach
    - Student crowd/team member
    - External code
  - When
    - Current resources/approaches arent productive or available
    - Need out of the box help with idea, code or members
- **Remove**
  - What
    - Ideas/approaches
    - Student crowd/teams
    - Project code
  - When
    - Not required, causing the project to slow down to move in wrong direction or for being unproductive

- **Rewire**
  - What
    - Student crowd/teams
    - Project code
  - When
    - Team member productive in an already productive team (or with specific skills, can be more useful in a potentially productive team.
    - If a piece of code developed by one unit of team seems to be more useful for other teams.
- **Compete**
  - What
    - Student crowd/teams
  - When
    - To foster growth of creativity (ideas) and deliveries (productivity)
- **Reverse**
  - What
    - Ideas/approaches
    - Student crowd/teams allocation
    - Project code
  - When
    - The current approach/process is less rewarding than the previous one, and its safe to reverse the approach rather trying new operators given current time and other situations

The first version of the algorithm consists of four primary operators, namely: *merge, split, add and remove*; and secondary operators, namely: *rewire, compete and reverse*. Using a combination of these operators, mentors (or organizers) can direct crowd and research process to maximize productivity under given resource constraints. The conditions under "when" will determine whether crowd/team needs to continue to merge, or split (stay parallel) or rewire etc. To measure the efficacy of the algorithm, weekly feedback survey can be conducted (Miller et al. 2014), complemented by peer-review or peer-grading techniques to track the progress of the project and/or student crowd. Finally, the quality of publishable results or paper, and its acceptance at a top-tier venue will determine the validity of the algorithmic process.

Preliminary Findings and Proposed Experiments

I am co-advised on this project by Prof. James Davis at UC Santa Cruz and Prof. Michael Bernstein at Stanford University. This project is already in progress, and was initiated in the Winter of 2014. So far, weve run two pilot studies in Winter and Spring academic quarters at UC Santa Cruz/Stanford. Moving further, encouraged by the findings from pilot studies, I plan to run this experiment in Fall 2014. Ill divide this section into the description and findings from pilot studies, followed by proposed experiment details.

The proposed experiment details can be divided into following subsections:

Pilot Experiment I

Pilot experiment I was a controlled experiment, conducted in a computer graphics class for graduate students in the Winter quarter at UC Santa Cruz. The aim of the experiment was to produce a crowdsourced publishable paper. In the beginning of the class, professor proposed three research problems to work on, and the class was divided into three major groups. The process involved students selecting the research problem, working on the implementation, generating results and writing sections of the paper in parallel. The students contributed to write sections in order of Related Work, Implementation, Methods and Results and finally Introduction and Conclusion. Using peer-grading systems, the best sections and results from each student were selected and made it to the paper.

This process resulted in three papers, however, unstructured. On the basis of student engagement and interest, the paper with maximum activity was selected to be worked upon further. To optimize the flow of the selected paper, a paper-a-thon was conducted post quarter, and after further refinement, paper was pre-peer reviewed twice by a group of four graduate students before sending to a relevant IEEE conference (ICIMu 2014). The results are awaited.

During this process, we learned that handling a project from idea inception to paper writing is a non-trivial task, and tends to produce partially unstructured paper. Post paper-a-thon efforts were required to improve the quality of the paper. We realized that we would want to narrow down the scope of this project to focus on producing publishable results, rather writing of the paper. We also learned that running an experiment in class has its own constraints, where most of the students are motivated for grades rather than doing novel research.

Pilot Experiment II

Based on the learning from pilot experiment 1, we conducted second round of controlled experiment in an undergraduate class in Spring quarter at UC Santa Cruz. To implement our learning, we focussed on the research process part instead of paper writing; and recruited students who were interested in research, and didnt involve students in the class who werent.

In collaboration with the Stanford Computer Vision Group, we proposed an open-ended research problem to the selected group of students. The project is in progress at Stanford, with few versions live. Students were blindfolded from the approaches used at Stanford, and encouraged to come up with their own ideas; which would later be compared against methods developed at Stanford. The process involved students to break up into four teams and working towards a common problem using their unique approach. Teams worked through a series of research phases such as: brainstorming, paper-pencil prototyping, development and user-evaluation. To boost productivity across these stages, student crowd and process were managed by Stanford and UC Santa Cruz researchers using algorithmic process and operators like split and merge. Using peer-grading systems, the best approach per-phase determined researchers feed-

back to all the participating teams, therefore, also encouraging others to iteratively improve their contributions.

The process though less organized, and used only two of currently proposed operators, proved to be effective. By the end of the quarter, solution proposed by the student crowd exceeded the ones developed at Stanford - qualitatively and quantitatively. The process learning leads us to run the proposed experiment. We learned that the algorithmic approach to manage student crowd and research process is effective. However, it brought up its constraints as well; therefore, I plan to add six more operators to take care of a variety of cases and situations. We also learned that though controlled experiment is useful, applying algorithmic process on student crowd in a class set up isnt quite feasible - students in a class tend to be inflexible upon frequent operator usage.

Overall, the two pilot experiments produced encouraging results and helped us focus on the caveats in future experiments.

Proposed Experiment

For the last two quarters, I have been running pilot experiments to understand an optimal way to crowdsource the research process. Based on my learnings, I propose an experiment which attempts to connect researchers with student crowd, and allow them to manage research process and crowd using a set of operators as part of the algorithmic approach.

Student recruiting As part of the experiment, I plan to recruit university students through a call for project - asking them to participate and get a chance to work with researchers at UC Santa Cruz and Stanford University. Due to low exposure of academic research in India, I plan to recruit students from Indian universities. So far, Ive convinced professors at Indian Institute of Technology (IIT Delhi and IIT Hyderabad) and Jaypee Institutes - a collection of public and private universities in India to help with the call for project. I plan to recruit about fifty students across these campuses, where students can create a team of up to five students. Though it might be hard to manage, I believe that due to course and other commitments, we might encounter a proportionate dropout of students. Interested students will sign up with their skill and time commitment information to participate.

The project With crowdsourcing research projects, the possibilities are virtually limitless. One can conduct ICT4D (Unwin and Water 2008) related research while sitting in the US, while some can run experiments which requires presence, without being present - through student crowd. Though the brainstorming about the project is still in progress, one of the potential research problem candidates is the following:

To identify an optimal combination of human computation and state-of-the-art computer vision algorithms to maximize image recognition accuracy relative to time and dollars invested. On the lines of Soylent (Bernstein et al. 2010) like interactive hybrid systems, student crowd will be encouraged to develop systems and measure accuracy against time and dollars and generate graphs to depict the same. The project is being considered due to its impact, few technical prerequisites, and easy quantitative evaluation of success.

Running the experiment and evaluation I plan to be running and organizing the experiment, and applying various operators on student crowd and research process as and when situation changes, from phase to phase. Factors determining usage of operators can depend on researchers feedback or peer review/grading metrics/results. I believe that the project might last for around a quarter to complete. To minimize researcher's time resource, students will get to meet online once per week, for about an hour. The agenda of the meeting would be to question any concerns, and to get feedback or direction to make further progress. To prevent any conflict, students will be told about the nature of the process, as flexibility to adapt as per operational changes is necessary. Specific steps will be taken to manage students dropouts.

The evaluation of the project will consist of getting weekly survey feedback from student crowd (Miller et al. 2014), peer-grading results, researchers feedback, and quality of graphs generated from the project under development. The project and this approach of crowdsourcing the research process would be a success if the results produced at the end of project is publishable at a top-tier venue.

Research Issues and Challenges

I believe that this project is trying to solve a non-trivial but high impact problem. The project delves in the areas of research and education, fostering the state of academic research via crowdsourcing. The project is not about implementing a software, but validating the algorithmic approach to manage crowd and research process. Running the experiment in the class had its own constraints, while running it as a call for project has its own. There are a lot of unknowns, and the scope of project still seems wide. Some of the specific research issues which I would like to discuss would be:

- Managing the student crowd remotely would be quite challenging, what processes or techniques can be used to streamline the task.
- Apart from the evaluation techniques and metrics mentioned above, what else should be measured to determine success.
- The algorithm proposed includes eight operators at the moment. Though it handles most of the scenarios anticipated to boost productivity. The practical implementation might be challenging.

  Though I have initial thoughts about handling these issues, brainstorming and getting expert feedback would help in addressing these challenges better.

References

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

Cranshaw, J., and Kittur, A. 2011. The polymath project: lessons from a successful online collaboration in mathematics. In *CHI*.

Dean, J., and Ghemawat, S. 2008. Mapreduce: simplified data processing on large clusters. In *Communications of the ACM*.

Dow, S. P.; Glassco, A.; Kass, J.; Schwarz, M.; Schwartz, D. L.; and Klemmer, S. R. 2010. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Trans. Comput.-Hum. Interact*.

2013. A formula for academic papers by jaime teevan. `http://slowsearching.blogspot.com/2013/11/a-formula-for-academic-papers.html`.

2014. Ieee international conference on information technology and multimedia at uniten (icimu 2014). `http://icimu.uniten.edu.my/index.php/about`.

Kim, J.; Cheng, J.; and Bernstein, M. S. 2014. Ensemble: Exploring complementary strengths of leaders and crowds in creative collaboration. In *CSCW*.

Kittur, A.; Smus, B.; Khamkar, S.; and Kraut, R. E. 2011. Soylent: A word processor with a crowd inside. In *UIST*.

2014. Leadgenius. `http://leadgenius.com`.

Miller, R. C.; Zhang, H.; Gilbert, E.; and Gerber., E. 2014. Pair research: matching people for collaboration, learning, and productivity. In *CSCW*.

2013. Is massive open online research the next frontier for education? `http://ucsdnews.ucsd.edu/pressrelease/is_massive_open_online_research_the_next_wi_for_education`.

Tomlinson, B.; Ross, J.; Andre, P.; Baumer, E.; Patterson, D.; Corneli, J.; and et. al. 2012. Massively distributed authorship of academic papers. In *CHI EA*.

Unwin, T., and Water, V. 2008. *ICT4D: Information and Communication Technology for Development*. Cambridge University Press.

Yu, L.; Kittur, A.; and Kraut, R. E. 2014. Distributed analogical idea generation: inventing with crowds. In *CHI*.