High-Speed Idea Filtering with the Bag of Lemons

Mark Klein, Ana Cristina Bicharra Garcia

***Abstract* - Open innovation platforms (web sites where crowds post ideas in a shared space) enable us to elicit huge volumes of potentially valuable solutions for problems we care about, but identifying the best ideas in these collections can be prohibitively expensive and time-consuming. This paper presents an approach, called the "bag of lemons", which enables crowd to filter ideas with accuracy superior to conventional (Likert scale) rating approaches, but in only a fraction of the time. The key insight behind this approach is that crowds are much better at eliminating bad ideas than at identifying good ones.**

***Index terms* – crowd-based, idea filtering, open innovation**

I. INTRODUCTION

OPEN innovation platforms (web sites where crowds post ideas in a shared space) enable us to elicit huge volumes of potentially valuable solutions for problems we care about, but identifying the best ideas in these collections can be prohibitively expensive and time-consuming (Riedl et al., 2010) (Schulze et al., 2012) (Westerski et al., 2013) (Blohm et al., 2011) (Bjelland and Wood, 2008) (Di Gangi and Wasko, 2009).

In response to this, organizations have turned to crowds to not just generate ideas, but also filter them, so only the best ideas need be considered by the decision makers. It has in fact been shown that crowds, under the right circumstances, can solve such classification problems with accuracy equal to or even better than that of experts (Surowiecki, 2005). This has been no panacea, however. Existing filtering approaches, when faced with large idea corpuses, tend to fare poorly in terms of accuracy, and can make unrealistic demands on crowd participants in terms of time and cognitive complexity (see http://ssrn.com/abstract=2501787 for a critical review of existing idea-filtering techniques).

This paper presents an approach, called the "bag of lemons", which enables crowds to filter ideas with accuracy greater than conventional (Likert scale) rating approaches, but in only a fraction of the time. The key insight behind this approach is that crowds are better at *eliminating bad ideas* than at *identifying good ones*. In the remainder of this paper, we will describe the approach, our experimental evaluation, and the lessons learned from the evaluations.

II. APPROACH: MULTI-VOTING WITH INCENTIVES

Our approach is simple. Raters are provided with the list of candidate ideas, as well as a clear description of the selection criteria. They are then given a limited number of "votes", and asked to allocate them to ideas based on whether or not they believe they represent top candidates for the decision makers. The more confident they feel about a judgment, the more votes they can allocate to that idea (within the limits of the overall vote budget). Raters are given financial incentives for allocating votes accurately. Ideas can then be filtered based on the number of votes each idea received. This approach is potentially attractive, we believe, because:

- *incentive alignment*: crowd participants are given incentives to make idea evaluations that align with those of the decision makers.
- *time demands*: rather than asking participants to rate all the ideas, they need only identify the small subset that they think are most (or least, see below) likely to be selected by the decision makers, rather than having to figure out the correct rating for *all* the ideas.
- *cognitive complexity*: participants are not required to deal with the cognitive overhead of trading and monitoring stock prices (as in idea prediction markets). In addition, as we will discuss below, the trick of asking users to assign votes to the *worse* ideas (rather than the best ones) further simplifies the evaluation process.

Our hypothesis, therefore, was that our multi-voting approach will allow crowds to achieve at least comparable levels of accuracy in filtering idea sets, while requiring less rater time, than conventional rating techniques.

III. EXPERIMENT DESIGN

To evaluate this, we engaged past and current members of a university R&D lab in identifying the most promising entries from a list of 48 ideas concerning how to increase productivity in the lab. The lab members were divided into three demographically matched groups of roughly 20 members each, each group using a different filtering approach:

- *Likert:* Participants were asked to rate each idea using a 5-point Likert scale, ranging from 1 (poor) to 5 (excellent) (Likert, 1932)
- *Bag of stars (BOS):* Participants were asked to distribute a budget of 10 "stars" to the ideas they felt were *most likely* to excellent.
- *Bag of Lemons (BOL):* Participants were asked distribute a budget of 10 "lemons" to the ideas they felt were *least likely* to be excellent.

The ideas were evaluated, up-front, by an expert committee, and participants were given financial incentives for accurately identifying the 19 ideas that were considered good or excellent

Mark Klein is a Principal Research Scientist in the Center for Collective Intelligence at the Massachusetts of Technology, as well as a visiting researcher at the University of Zurich and a visiting professor at the Nagoya Institute of Technology http://cci.mit.edu/klein/

Ana Cristina Bicharra Garcia is a full professor in the Computer Science Department at Fluminense Federal University (UFF) in Brazil. http://www.addlabs.uff.br/Novo_Site_ADDLabs/index.php/en/about

by at least three members of the expert committee. All the idea filtering engagements took place in parallel, participants could not see each other's ratings, and were asked to not discuss their evaluations with each other during the experiment, to help assure the rater independence that is required for accurate crowd classification (Ladha, 1992). All user interactions with the system were recorded and time-stamped.

IV. EVALUATION RESULTS

We used a standard technique known as ROC curves (Fawcett, 2004) to assess the accuracy of the idea filtering methods. ROC curves plot the true positive rate vs. the false positive rate for a filter. The area under the ROC curves is then a measure of accuracy: a perfectly accurate idea filter would have an area of 1.0, while a random selection filter would have an area of 0.5.

The accuracy scores for the three idea filtering conditions were as follows:

Condition	Accuracy
BOL	0.89 +/- 0.04
Likert	0.74 +/- 0.03
BOS	0.62 +/- 0.03

BOL had the highest accuracy, followed by Likert and then BOS. All conditions performed better than a random filter (which would have an accuracy of 0.5), and all these differences were statistically significant at $p<0.05$.

The average amount of time the participants spent, in minutes, doing the ratings in each idea filtering condition were as follows:

Condition	Per-Rater Time/minutes
BOL	24 +/- 12
Likert	75 +/- 20
BOS	25 +/- 28

BOS and BOL required roughly 1/3rd the rater time of the Likert approach ($p<0.05$). The difference between BOS and BOL was not statistically significant.

Our data allows us to reach the following conclusions:

- The bag of lemons (BOL) approach provided substantially (about 33%) greater idea filtering accuracy than the conventional Likert approach, while requiring only about one third of the rater time.
- Our crowds were much (about 60%) more accurate *at eliminating bad ideas* (BOL) than selecting good ones (BOS).

Our hypothesis (that our approach will achieve at least comparable idea filtering accuracy as Likert rating, while requiring less rater time) was thus validated for the Bag of Lemons, but not for the Bag of Stars.

V. LESSONS LEARNED

How can we understand these results? We believe that the key insight is that identifying the *best* ideas requires finding ideas that are exceptional with respect to *all* relevant criteria (e.g. feasibility, value, and cost). This can be time-consuming and, in addition, may force raters to make judgments that they are not well-qualified to make. A rater, for example, may have a good sense of the potential benefits of an idea, but not of how costly it would be to implement. The Bag of Lemons approach, by contrast, tries to find the *worst* ideas, and this only requires that people identify ideas that are clearly deficient with respect to *one criterion*, since that is all it takes to eliminate an idea from consideration. The incentives and limited vote budget, in addition, encourage raters to focus only on the ideas they feel they can evaluate quickly and well. As long as the rater community is diverse enough so that every criterion has at least some raters who can evaluate it, the bag of lemons can achieve greater idea filtering accuracy than any member could achieve on his/her own, while reducing rating time as compared to techniques that require raters to evaluate with respect to all the criteria.

This work represents, we believe, a novel and important contribution to the literature on idea filtering for open innovation systems. While other efforts have used multi-voting for idea filtering (Bao et al., 2011), or provided incentives for idea filtering accuracy (e.g. in prediction markets), we are aware of no previous work that combines these concepts, or that is based on identifying the worse, rather than the best, ideas.

VI. ACKNOWLEDGMENTS

M. Klein's participation in this work was supported by grant 6611188 from the European Union Seventh Framework (FP7) Program - the CATALYST project. A.C.B. Garcia gratefully acknowledges funding support from the Brazilian government research agency (CAPES).

VII. REFERENCES

[1] Bao, J., Sakamoto, Y., & Nickerson, J. V. (2011). *Evaluating design solutions using crowds*. Proceedings of the Seventeenth Americas Conference on Information Systems,.

[2] Bjelland, O. M., & Wood, R. C. (2008). An Inside View of IBM's 'Innovation Jam'. *Sloan Management Review*, *50(1)*(1).

[3] Blohm, I., Bretschneider, U., Leimeister, J. M., & Krcmar, H. (2011). Does collaboration among participants lead to better ideas in IT-based idea competitions? An empirical investigation. *International Journal of Networking and Virtual Organisations*, *9(2)*(2), 106-122.

[4] Di Gangi, P. M., & Wasko, M. (2009). Steal my idea! Organizational adoption of user innovations from a user innovation community: A case study of Dell IdeaStorm. *Decision Support Systems*, *48(1)*(1), 303-312.

[5] Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, *31*, 1-38.

[6] Ladha, K. K. (1992). The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 617-634.

[7] Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, *140*, 1-55.

[8] Riedl, C., Blohm, I., Leimeister, J. M., & Krcmar, H. (2010). *Rating scales for collective intelligence in innovation communities: Why quick and easy decision making does not get it right*. Proceedings of the International Conference on Information Systems.

[9] Schulze, T., Indulska, M., Geiger, D., & Korthaus, A. (2012). Idea assessment in open innovation: A state of practice.

[10] Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor.

[11] Westerski, A., Dalamagas, T., & Iglesias, C. A. (2013). Classifying and comparing community innovation in Idea Management Systems. *Decision Support Systems*, *54(3)*(3), 1316-1326.