Modeling Problem Difficulty and Expertise in StackOverflow

Benjamin V. Hanrahan
Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA, 94304, USA
hanrahan@parc.com

Gregorio Convertino
Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA, 94304, USA
convertino@parc.com
Xerox Research Centre Europe
6 Chemin Maupertuis
38240 Meylan, France
gregorio.convertino@xrce.xerox.com

Les Nelson
Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA, 94304, USA
nelson@parc.com

CSCW 2012, February 11–15, 2012, Seattle, Washington.
ACM 978-1-4503-1051-2/12/02.

Abstract

Supporting expert communities is becoming a 'must-have' capability whenever users are helping each other solve problems. Examples of these expert communities abound in the form of enthusiast communities, both inside and outside of organizations. In order to achieve success, these systems have to connect several different actors. In this paper we inform the design of these ***Hybrid Intelligence Systems*** through the investigation of StackOverflow. Our focus in this paper is to develop indicators for hard problems and experts. The long-term goal of our study is to examine how complex problems are handled and dispatched across multiple experts. We outline implications for modeling these attributes and how they might inform better design in the future.

Author Keywords

Community of experts, CSCW, collective intelligence, crowdsourcing

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Design, Measurement

Introduction

Enthusiast crowds are a valuable, scarce resource. The scarcity of qualified experts increases alongside problem difficulty. Due to the scarcity of qualified experts, our goal is to effectively utilize the differing expertise of enthusiast crowd members. That is, for the numerous, simple problems we should engage the numerous, low-expertise crowd members. Likewise, for the few, complex problems we should engage the few, high-expertise members of the crowd.

In order to design such systems, accurate indicators are required for two aspects, problem difficulty and user expertise. To develop these indicators, we have chosen to investigate the StackOverflow community [5]. StackOverflow is a successful, active community of software developers that provide answers to each other's questions. In our dataset we had 1.9 million questions, 4.2 million answers, 6.9 million comments, and 640 thousand users. Through the analysis of this community we expect to be able to discover the aspects of a problem that make it difficult, the aspects of a user that make them a qualified expert, and the interactions between the two.

In this paper we review related work, outline the methodology that we are using to analyze the StackOverflow community, provide preliminary results for the study, and outline future work.

Related Work

A classic success case of a help-based community in an organization is Eureka. Eureka is a system for practical knowledge-sharing deployed twenty years ago to facilitate sharing among Xerox technicians. These technicians repair devices as a service for Xerox customers [2]. Another classic example of a help-based community is the community around Ackerman's Answer Garden system [1]. Answer Garden, allowed users to find answers to commonly asked questions and gain access to experts in the organization.

Several studies of Q&A and technical support communities have measured and modeled the expertise of community members. Zhang, Ackerman, and Adamic [4] investigated the Java Forum, a help-seeking community for Java programmers. They characterized the expertise network and compared alternative algorithms for ranking expertise. They found that structural information and social network-based algorithms can be used for evaluating an expertise network. A recent study that analyzed StackOverflow [3], found that the community is lively and successful. In fact, over 92% of questions are answered, in a median time of 11 minutes. Note however that this study did not focus on problem difficulty, or modeling relationship among difficulty and user attributes.

Research Method

StackOverflow publishes their data under the creative commons license. This enabled us to rebuild the database from the published xml files and analyze the raw data from the community. The two aspects that we report in this paper were *question difficulty* and *user expertise*.

StackOverflow has several interesting attributes that particularly suit the dataset to our needs. First, the experts on StackOverflow are often professional programmers that dispense valuable advice and information. Second, there are several clear usage metaphors that let us put bounds on whether and when a question is considered answered by the poster. Third, the detailed data at our disposal lets us extract interesting events such as when a question is edited or when an answer is commented on.

Question Difficulty

Through inspection of the data, we decided to use the duration between the when a question was first posted and when an answer was accepted by the poster as a proxy for difficulty. This enabled us to correlate different measures with this duration.

We then extracted all of the events for each question; our aim was to construct a formal language that described the activity around a question. The grammar we used to construct this language was: *Actor Type – Action – Object*. We extracted the following actor types: *Questioner*, *Accepted Answerer*, *Unaccepted Answerer*, and *Someone*. Respectively, these are the user who asked the question, the user who posted the accepted answer, users who posted answers that were not accepted, and anyone else that participated in the question. We extracted the following actions: *Posting*, *Commenting*, *Editing*, *Answering*, *Accepting*, *Voting Up*, and *Voting Down*. We then extracted which object was being operated on, either a *Question* or an *Answer*. Through this language we were able to categorize a diverse set of events. For example, the questioner commenting on an answer, the accepted answerer editing the question, or someone voting up for an answer.

After constructing these events with our grammar, we then binned the occurrences of each event per question. These different bins were then correlated with the aforementioned duration.

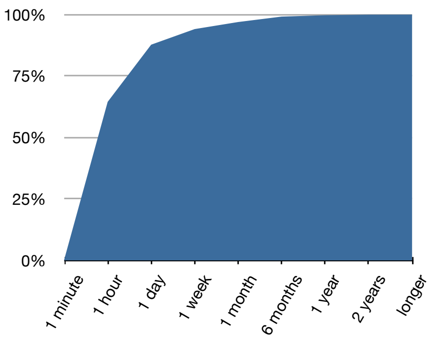

Figure 1. Percentage of questions answer at a given time.

User Expertise

In order to investigate the interactions between question difficulty and expertise we used three different measures. First, we used the reputation as displayed and calculated by StackOverflow. Second, we used the Z-score as reported by Zhang et al [4]. Lastly, we used the average delta between up and down votes for all previously submitted answers.

Using these measures we calculated the expertise of the different roles involved with each question, i.e. *Questioner* expertise, *Accepted Answerer* expertise, etc. We then correlated these expertise scores with question duration. Our hypothesis was that questions that took longer to answer required a higher degree of expertise to finally answer.

Preliminary Results

Characterizing Problem Difficulty

Since we used the duration between question and accepted answer postings as a proxy for difficulty we provide the following histogram detailing the percentage of questions answered after a given period of time, see Figure 1.

Initially we measured the Pearson's product-momentum correlation coefficient between all of our binned events and the duration of the question. This did not reveal any strong correlations. After inspection of the data we decided to measure the Spearman's rank correlation coefficient to decrease the sensitivity to outliers. This test yielded more interesting results. Supporting one of our anecdotal observations that sometimes especially difficult questions are in fact answered by the questioner, $\rho=0.31048$ with $p < 0.$

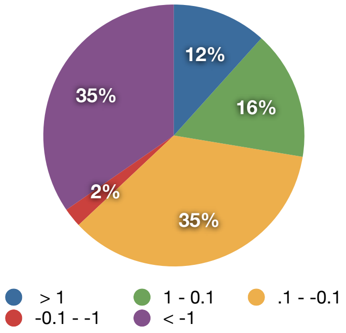

Figure 2. Percentage of users with Z-score in indicated range (644,000 total).

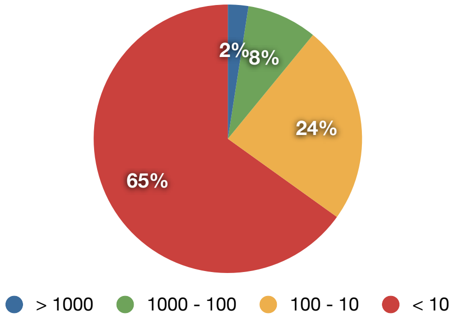

Figure 3. Percentage of users with reputation score in indicated range (644,000 total).

We are continuing our investigation into which events and sequences indicate problem difficulty.

Characterizing Expertise

As a first step in examining the reputation and z-scores of the StackOverflow community, we produced a pie chart to show the distribution of users in regards to these scores. The percentages of users that have Z-scores in the indicated ranges are provided in Figure 2. Users with a positive Z-score are answering more questions than they are asking, while users with a negative score are asking more questions than they are answering. In Figure 3, we provide similar percentages in regards to the reputation of users as scored by StackOverflow. The reputation score on StackOverflow captures the activity of a user on the site; points are awarded for answering questions and performing various maintenance tasks. We then examined the Pearson's product-momentum correlation coefficient for these variables and found that they had a strong correlation, $r = .60363$ with $p < 0.$

We have yet to find a strong, significant relationship between the duration of a question and the average expertise of the involved users. Although, we are expertise that take into account the infrequent activities of high-expertise users. Particularly, since both the reputation and Z-score heavily favor more active users.

Discussion and Future Work

The roadmap for this research is to further develop the measuring and modeling of problem difficulty and expertise. These will enable the exploration of how problem difficulty relates to the expertise of the crowd, and whether or not different social network structures form around different questions. Beyond completing our analysis on StackOverflow we plan to replicate our analyses on Eureka, another clearly successful community of experts. In both cases we plan to set up experiments that manipulate variables such as problem difficult and observe the outcomes. The results of these experiments can be used to validate or tune our models.

As so far as leveraging the lessons we learn, we plan to build a prototype of an hybrid intelligence system that models key properties of the problem and the human agents and leverages the resulting models to optimally delegate tasks across a diverse pool of intelligence agents (human and artificial).

Acknowledgements

We would like to thank the StackOverflow community for making their data publicly available.

References

[1] Ackerman and McDonald. Answer Garden 2: merging organizational memory with collaborative help. In Proceedings of CSCW '96. ACM, New York, NY, USA, 97-105. 1996.

[2] Bobrow, D.G. and Whalen, J. The Eureka Story: Community Knowledge Sharing in Practice. Reflections, 4 (2). 2002.

[3] Mamykina, Manoim, Mittal, Hripcsak, and Hartmann. Design Lessons from the Fastest Q&A Site in the West. In CHI '11. ACM, New York, NY, USA, 2857-2866. 2011.

[4] Zhang, Ackerman, and Adamic. Expertise networks in online communities: structure and algorithms. In WWW '07. ACM, New York, NY, USA, 221-230. 2007.

[5] http://stackoverflow.com