Discovering Elite Users in Question and Answering Communities

Cheng-Yue, Royce
rchengyue@cs.stanford.edu

Hsu, Richard
rhsu@cs.stanford.edu

Stevens, Nicholas
nstevens@cs.stanford.edu

Abstract

Question and Answering (Q&A) communities depend on a set of users who have mastery of the topics being discussed and also actively respond to questions. We will refer to these users as "elite users". Identifying these elite users allows general users to identify credible sources. It can also help in allowing community designers to direct unanswered questions to these elite users for a higher chance of a response and credibility. The goal of our work is to investigate the activity and behavior of users in a particular Q&A community and discover whether or not it is possible to predict elite users based on early signals.

Keywords.
Question-answering, reputation, value prediction.

General Terms.
Experimentation, human factor, measurement.

1 Introduction

Online education has been growing rapidly over the last few years. Many of these systems use some aspect of a question and answering forum where anyone can ask a question and anyone in the community can respond. These communities provide a completely community-driven knowledge portal. However, much of this process relies on the fact that a subset of users in the community have some mastery of the topics; otherwise, the questions would rarely be answered. We refer to these users as "elite users" and are those who contribute actively and are reliable sources. These users form the most important aspect of these forums and provide the necessary driving force to make such systems successful.

Many websites reward such users with some form of reputation that can be granted through votes and rewards for responding, thereby allowing users to identify those with mastery of the topics. Unfortunately, gaining such status takes time and hides the proficiency of an early user.

We believe that identifying elite users will be beneficial because it provides community designers the ability to direct unanswered questions to these elite users for highly reliable and credible responses. It can also help by allowing community designers to grant special privileges to these users earlier on in their career, thus allowing them to better utilize their expertises. To assist in the identification of these elite users, we explore a particular community called Stack Overflow[1] to see if early attributes or activities associated with users can help distinguish elite users.

2 Related Works

There has been previous research involving question and answering communities that focused mainly on analyzing questions and answers and gaining insight about their properties [1, 5]. One particular study looked into predicting long-term value of a question in the Stack Overflow community [1]. We believe that, in conjunction with such work, being able to identify elite users can provide Q&A communities the ability to bring more attention to these users sooner. This would help the site match the most reputatable users with the constant influx of new questions and further improve upon the reliability and long-lasting value of questions and answers.

This study also provides a parallel evaluation for long-term value of an entity in these Q&A communities. A. Anderson, et al. primarily found a high indication of utilizing intrinsic properties of the questions to predict their long term value, which we believe has a similar parallel for users [1]. We believe that elite users have some common intrinsic properties that help distinguish them from regular users. Although identification of long-term value in users may require looking into larger time frames of data as inputs, there are many properties of users that could act as features to help predict their likelihood to become elite users in a year.

Besides evaluation of long-term value in questions in the communities, the identification of the "expert" set of users in these Q&A communities has been explored in the research community as well [3, 4]. Our work is different in that not only do we want to identify experts, but we also want to identify expert users who are highly active and can make the biggest impact. Additionally, we want to discover value from their behaviors that not only distinguish them in the community but also allow us to predict and identify elite users early on in their careers. Such work is similar to A. Pal, et al. as they argue that evolutionary data of users can be more effective at expert identification than the models that ignore evolution [2]. We want to complement this work with further experimentation on other features expressed by users on Stack Overflow and further discover attributes that separate elite users from regular users.

By continuing previous works in predicting entity values in Q&A communities, we want to show that it is possible to distinguish elite users early on and not have to rely on years of actions in order for them to gain credibility in the community. This is important because a past study showed that users who contributed a lot had greater influence than new users [7]. This intuitively makes sense because those already contributing continue to maintain their long-lasting value; however, new users who may be elite users are greatly undervalued.

3 Data Set Description

Stack Overflow employs a targeted model in both domain and question type encouraged on the site. As opposed to the myriad of popular general Q&A sites on the web today (Yahoo! Answers, Quora), Stack Overflow advertises itself as a programming only Q&A site. Furthermore, all questions posed on the site are meant to be looking for a single, 'best' answer. Subjective questions that have no hope for such a definitive response are usually weeded out.

This focus on both domain and question type is only successful thanks to the users. First, the subset of the population equipped to answer programming questions is small, and the value question askers acquire is largely based on having the right expert answer their question. Second, many of the top users serve as de facto moderators, removing questions that do not fit the mission of Stack Overflow and merging duplicates. To ensure this power is not abused, users are incrementally granted increased abilities based on their own reputation.

The possible and quantifiable actions that occur on Stack Overflow extend beyond simply asking and answering. Users can both comment on questions and answers, all forms of communication can be voted on, and users can favorite questions. Finally, any of the possible answers can be designated as the 'accepted answer' by the original asker. All of these attributes are used to present each question on the site: after displaying the question at the top, the answers and comments are stack ranked based on the number of up minus down votes, with the 'accepted answer' always presented first.

For our project, we are using a complete trace of the site that extends from the sites inception on July 31, 2008, to August 7, 2012. Some basic statistics of this dataset can be seen in Table 1. We are using MySQL to query the data and the Python library `sklearn` to

[1] http://www.stackoverflow.com

	Total	Other Statistics
Users	1,295,620	55.25% asked a question 38.60% answered a question
Questions	3,453,742	62.21% accepted answers
Answers	6,866,609	32.15% accepted
Votes	21,460,580	93.00% positive
Favorites	1,992,831	on 788,991 questions

Table 1: Stack Overflow's Statistics

Action	Author	Action Taker
Answer is upvoted	+10	+0
Answer is downvoted	-2	-1
Answer is accepted	+15	+2
Question is upvoted	+5	+0
Question is downvoted	-1	-2
Answer wins bounty	+Bounty	-Bounty
Answer marked as spam	-100	+0
Accepted suggested edit	+2	+0

Table 2: Stack Overflow's Reputation System

help with our prediction models. Since our predictions will be relating to evaluation of users being elite or not, we rely on the reputation system that Stack Overflow has developed based on the actions taken by each user (Table 2). We use this evaluation because it is the social evaluation of whether or not a user is of any value to the community and given enough time should be a good indicator of the value of the user to the community.

4 Analysis of Dataset

We began by looking at the dataset to learn more about the community and verify some of our possible intuitions about the network. We looked at some various properties and found the dynamic between questions, answers, and reputation to be rather interesting.

4.1 Reputation Before and After

We began by looking at cohorts of users who joined Stack Overflow in specific time frames. In other words, each cohort of users has a specific time window since the users' inceptions to interact with the Stack Overflow community. After exploring the data, we found that the correlation coefficient was 0.797 between the reputations of cohort of users who joined 3 months before September 7, 2011 versus their reputations roughly a year later on August 7, 2012. Given the plot of the users' reputations before and after in Figure 1 we can see that the effective loss in the top right quadrant is smaller than in the bottom right quadrant. This leads us to the intuition that getting higher reputation scores early on is a good predictor of separating out elite users later on.

4.2 Reputation's Social Evaluation

The reputation of Stack Overflow heavily favors those that answer questions. Based on the reputation system, authors of answers stand to gain more reputation than authors of questions (see Table 2). This shows the importance of answers because Question and Answering

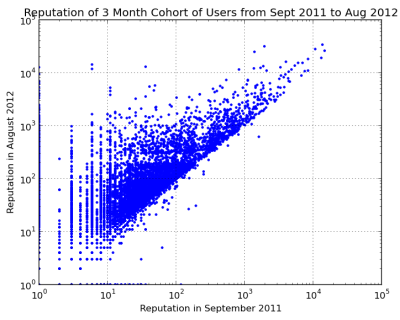

Figure 1: Reputation of a cohort at 3 months compared to a year later.

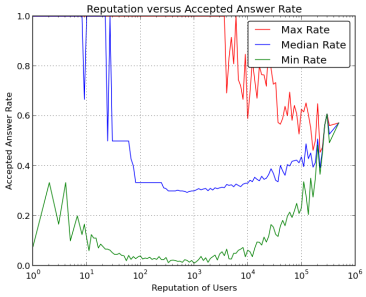

Figure 4: Reputation versus Accepted Answers

communities rely heavily on users who have mastered the topics to respond and answer questions. In Figure 2, we have plotted user reputations as of August 7, 2012 and the max, median and minimum number of questions and answers for a user to be a member of each cohort. For the maximum number of post types, the values are very similar to the general population of users, and the trend is that there is a similar number of answers and questions for all reputations below 10,000. After 10,000 reputation, we see a separation between the answer and question counts, signifying that top tiered users generally have more answers than questions. This coincides with our intuition of the built in reputation system, which uses reputation to evaluate the "eliteness" of a user over time. Similarly, this holds for the rest of the population as can be seen by the median and minimum graphs in Figure 2. We want to explore the idea of elite users, which we define as those who can largely contribute to the community through activity and trustworthiness. Since this intuition is consistent with the above reputation analysis, we will utilize reputation to help classify and validate our prediction model.

4.3 Average Views

We want to find the top tier of users who are active and credible. As a result, we explored the average view counts over a user's set of answers and accepted answers. In figure 3, we notice some interesting information. For low reputation users, we see a higher average view count of answers given than that of accepted answers. This occurs for the maximum, median, and minimum; however, as we reach 10,000 reputation, we see that these values begin to converge. This is similar to the threshold where the number of answers given and the number of questions given diverge for maximum number of posts in Figure 2. Consequently, an elite user's accepted answers and given answers are more than likely correlated as compared to a regular user who may have more answers but very few accepted ones. Finally, we look at the max average views for the maximum of these average views (far left graph in Figure 3). This shows that, even with high number of page views, a user is not necessarily going to gain large reputation. Thus, the number of answers and popularity of a post do not coincide with developing intuitions of necessarily having high contributions to the community. Specifically, this does not happen until a user's accepted answers gain more popularity. Even then, we notice that the accepted answer's max average views is staggeringly lower than that of answers given. We also see that the average page views decrease as the reputation of a user increases. As a result, even though a post is popular, it does not necessarily portray the same social measures as being credible or worthy.

4.4 Average Accepted Answers

Our final discussion of our analysis looks at the accepted answer rate of users versus their reputations. Overall, the maximum and median plots do not tell us much except that there are possibly many users who have a few questions that have been accepted. Reasonably, a new user can have his or her answer accepted, and, as the answer gains reputation, the user gains reputation. The most intruiging part of Figure 4 is the minimum accepted answer rate. Intuitively, a new user is usually motivated to join if they have a specific question to ask or answer to contribute. This leads us to think that their first action of either type should be of a higher quality. In this way, new users can have a decent answer rate if they can achieve few accepted

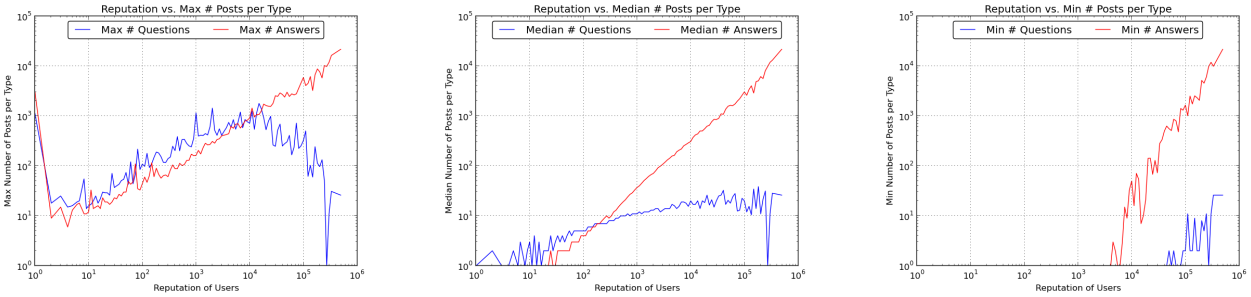

Figure 2: Reputation vs. Number of Posts on Log-Log scale. Number of Answers Given (red) and Number of Questions Asked (blue). Maximum number of posts given a reputation (left). Median number of posts given a reputation (middle). Minimum number of posts given a reputation (right).

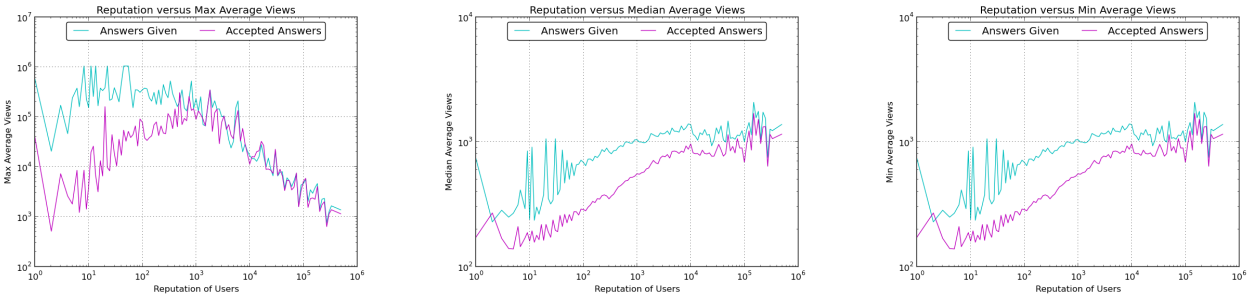

Figure 3: Reputation vs. Average Views on Answers or Accepted Answers. Answers Given(cyan) and Answers Accepted (magenta). Maximum average views (left). Median average views (middle). Minimum average views (right).

answers. However, we see that, as reputation increases, this minimum average accepted answers decreases to almost 0% and then begins to rise again at around 10,000 reputation (the transition point similar to those seen in previous analysis sections). At this point, we see that the accepted answer rate begins to increase to a point where it converges at almost 60% acceptance rate. This phenomenon indicates that top tiered users who have lots of reputation tend to have higher average acceptance rates for their answers. There are still many possibilities for a user to gain reputations; however, from analyzing these plots, we see that higher answer acceptance rates generally correspond to higher probabilities of elite users.

To conclude this section, we see that the characteristics of answers help describe whether or not a user will become elite. We aim to not only look at the user's characteristics, but also the answers users give and the types of questions they seem to target. By looking into a variety of features, we hope to capture a few essential signals that will help identify early elite users.

5 Methods

To predict whether or not users will be elite the following year, we decided to use two different classification algorithms: logistic regression and SVM with an RBF kernel (RBF-SVM). In general, logistic regression is the standard method for numerous classification tasks and tries to optimally find a linear decision boundary among the data. In the case that our data is not linearly separable, we also decided to use an SVM to generate more complex decision boundaries in the hopes of identifying possible nonlinear structures and achieving a higher performance. A common choice for this purpose is the RBF-SVM.

In addition to applying these algorithms with a full set of features, we also decided to have baselines using logistic regression and RBF-SVM on the following three features: reputation, upvotes, and downvotes. From our data anaylsis on early reputation values (Figure 1), we found that there is a high correlation between the early reputation of a user and the reputation of a user the following year. Because Stack Overflow uses upvotes and downvotes as main sources of reputation and direct evaluations of a user's questions and answers, we decided to add the number of upvotes and the number of downvotes to the baseline features. We used these baselines to determine whether or not our chosen models and predictions provide meaningful insights into the development of elite users.

To evaluate our performance, we used classification accuracy, area under the ROC curve (AUC), and F_1-scores. Although classification accuracy is a standard metric in determining the success of the classification algorithms, this measurement usually provides misleading results for highly skewed data. As a result, we performed undersampling on the data set to balance the class label distribution before reporting our evaluation results [8]. AUC measures the relationship between true positive rates and false positive rates and an F_1-score is defined as:

$$2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

In order to train our models, we calculated and used the following standardized features:

- **Characteristic of the User:** Number of questions asked, number of answers given, number of accepted answers, number of comments, number of personal favorites, amount of bounty, number of badges, number of user profile views, number of upvotes, number of downvotes, reputation.
- **Characteristics of All of a User's Questions:** Accumulated score, accumulated views, accumulated favorites, accumulated number of answers, accumulated length of questions, accumulated length of title of questions.
- **Characteristics of All of a User's Answers:** Accumulated score, accumulated number of comments, accumulated length of answers, accumulated score of the corresponding questions, accumulated views of the corresponding questions, accumulated favorites of the corresponding questions, accumulated number of other answers to corresponding questions, accumulated length of the corresponding questions, accumulated length of the titles of the corresponding questions.

	Baseline		Full Features	
	Logistic	**SVM**	**Logistic**	**SVM**
1 Week	0.69	0.68	0.64	0.63
1 Month	0.74	0.75	0.77	0.75
3 Months	0.83	0.83	0.84	0.83

Table 3: Classification accuracy for logistic regression and SVM for both baseline and full feature set.

	Baseline		Full Features	
	Logistic	**SVM**	**Logistic**	**SVM**
1 Week	0.70	0.70	0.70	0.69
1 Month	0.80	0.81	0.81	0.81
3 Months	0.88	0.88	0.89	0.89

Table 4: Area under the ROC curve for logistic regression and SVM for both baseline and full feature set.

	Baseline		Full Features	
	Logistic	**SVM**	**Logistic**	**SVM**
1 Week	0.60	0.57	0.56	0.54
1 Month	0.68	0.68	0.73	0.71
3 Months	0.81	0.81	0.83	0.81

Table 5: F-1 scores for logistic regression and SVM for both baseline and full feature set.

Our classification task is to predict whether or not users will become elite users based on what they have accomplished in a time frame since they joined Stack Overflow. To simplify our model, we defined an elite user to be in the top 10% percentile of a cohort based on reputation. While this is an arbitrary choice, we viewed any larger percentiles (i.e. top 25%) would lower the difficulty in achieving high evaluation scores in this prediction task and would make it too easy for a user to be classified as elite. Also, we believe only the top performing users should be characterized as those who contribute immensely to the community.

For our predictions, we decided to use the idea of cohorts. Specifically, we sample users at time frames of one week, one month, and three months since their inceptions and look at their features after this time frame. We use a user's feature vectors as examples and whether or not their reputation is in the top 10% of their cohorts the following year as labels. Specifically, we trained our classification algorithms on September 2010 and September 2011 data and tested our classification algorithms on September 2011 and August 2012 data.

The reason we do not use labels corresponding to the top 10% of the global population is that we want to take into account the discrepancies between the user start dates. Specifically, a user who joined in 2010 would have an additional year to accumulate more reputation than a user who joined in 2011. Using the idea of cohorts, we essentially normalize the data set to remove any extra advantages users have for joining earlier.

After running our classification algorithms on our baseline and full feature sets, we decided to use feature selection to determine a set of essential features and focus our analysis on these imperative features.

6 Results

Table 4 and 5 provide the results found through logistic regression and RBF-SVM on our baseline and full feature sets. From these results, we find that our prediction model for the one week cohort is considerably less accurate than that for the one month and three month cohorts. Intuitively, this result makes sense as it shows that the longer a user spends in the community, the better prediction we can make in their future community contribution level. In terms of the specific algorithm performances, we find that logistic regression provided similar or better results than RBF-SVM across all time frames for both feature sets. However, after users spend three months in the community, logistic regression performs better than RBF-SVM. We also notice that, although logistic regression performed similarly for the full feature set and the baseline feature set in terms of AUC scores and classification accuracies, logistic regression performed much better with the full feature set than with the baseline feature set in terms of F_1 scores. RBF-SVM with the baseline feature set performed better than with the full feature set for the one week cohort; however, RBF-SVM with the full feature set consistently had higher accuracies for the other time frames.

After running feature selection, we find that there are nine essential features for each time frame. Performing logistic regression and RBF-SVM on these nine essential features shows about a 0.01 decrease in classification accuracies, 0.01 decrease on AUC scores, and 0.03 decrease on F_1 scores from using the full feature set on all time frames. However, if we remove a feature from this set of essential features, our evaluation metrics take a considerable performance hit of around 0.05 decrease in classification accuracies, 0.05 decrease on AUC scores, and 0.1 decrease on F_1 scores.

To further analyze these top nine core features, we decided to look at the weights vector generated by logistic regression. The relative importance of the essential features with respect to each time frame is shown in Table 6. By analyzing the weights generated from our nine essential features, we find numerous intriguing points. First, we notice that there are four consistent features across all the time frames: accumulated favorites of the corresponding question to a user's answer, accumulated number of comments on the user's answer, number of accepted answers, and reputation.

For a brand new user, reputation is not a strong predictor of whether or not a specific user will eventually become elite. This result is consistent with our data analysis section earlier. Specifically, once we allow users a month to contribute to the Stack Overflow community, we begin to see that reputation becomes a stronger predictor of whether or not they will be an elite user the following year. This phenomenon follows closely to Figure 2, in which we begin to see divergences once a user has begun accumulating more reputation. Another similar feature is that of number of accepted answers. Elite users are very important to a community with respect to their active contributions and their credible backgrounds. As we allow elite users more time to answer questions in the community, we see that the number of accepted answers becomes a stronger predictor for an elite user. Both reputation and number of accepted answers are intuitive features that we hope would shine in the prediction model.

The accumulated number of comments for a specific user's answers and the accumulated favorites of the corresponding questions of a specific user's answers are two other features that contribute to the prediction of an elite user. We notice that these two features are related in the sense that they involve community engagement. Specifically, if a user answers an important or thought provoking question in the community, it is likely that the user's answers will generate discussion among other users, causing a large number of comments for the user's answers. Similarly, we find that, as time progresses, another common feature of elite users is the general decline of importance in the questions they are responding to, which can be estimated by the number of favorites on their corresponding questions. We see that, in the beginning, elite users tend to answer highly favorited questions to establish themselves. However, as they spend more time on the site, they begin to answer questions that are specific to the individual and are not necessarily high in favorites. In this way, by answering a variety of questions, elite users welcome other users to ask questions regardless of the questioner's status. From this analysis, we find that elite users not only contribute meaningful answers, but they also involve and welcome other members of the community in their questions and answers.

Apart from these four common features in all the time frames, there are four other features that play an important part in determining whether or not a user will eventually become elite: accumulated score of the corresponding questions, the total number of comments the user has made, accumulated length of their answers, and the number of downvotes. Interestingly, we find that the accumulated scores of questions a user answers has a negative weight in the one week time frame and a small weight in the three month time frame. This result means that elite users do not only answer questions with large number of upvotes early on, but they answer a variety of questions, including those that may not be as popular or heavily upvoted. The number of comments and the accumulated length of the user's answers are two features that relate back to the idea that elite users not only provide thoughtful insights to questions, but they also generate community engagement among their peers. The number of downvotes is also expected to have a negative weight; however, this is intuitively a

Feature	Coefficients 1 Week	1 Month	3 Month
Accumulated favorites of the corresponding questions	+1.757	+0.266	-0.011
Accumulated number of comments for user's answers	+2.613	+1.297	+2.405
Number of accepted answers	+0.174	+0.761	+1.364
Reputation	+0.676	+3.422	+7.533
Accumulated score of the corresponding questions	-0.508	–	+0.009
Number of comments	+0.325	–	+0.409
Accumulated length user's question	–	+0.496	+0.296
Number of downvotes	–	-0.015	-0.121
Accumulated number of answers for a user's questions	+0.051	–	–
Accumulated score of the answers	+3.132	–	–
Number of answers given	-0.009	–	–
Accumulated number of other answers to corresponding questions	–	+0.711	–
Accumulated views of corresponding question of user's answer	–	-0.199	–
Accumulated question favorite count	–	+0.051	–
Number of personal favorites	–	–	+0.375

Table 6: Top coefficients using logistic regression with nine essential features on the 1 week, 1 month, and 3 month cohorts.

high-precision, low-recall feature. Specifically, if a user is downvoted many times, the user is most likely not an elite user; however, having low number of downvotes does not necessarily indicate that the user will eventually become an elite user.

7 Conclusion

Our results show that we can reasonably predict whether or not users will become elite users in the future from their initial behaviors. Although our prediction model does not improve much over our baseline analysis, we gained a great amount of insight on the interaction between the current evaluation system of reputation and what we would want reputation to represent. We see that early reputation is a great indicator of whether or not a user will be elite; however, we also know that other interactions and behaviors also provide clues. By performing a classification task to forecast elite users, we find that our intuitions are on the right track and we are able to devise a prediction model with appropriate sets of features.

We initially approached this as an implementation exercise, but ended up learning far more. First, examining the data and deciding on an approach proved to be very important. We were lucky to have a large dataset to work with, but its size made initial data analysis before exploring possible features paramount. Next, we learned a lot about the varying methods of both producing the necessary features and running our algorithms efficiently. When dealing with such large cohorts of users who each take thousands of actions, the efficiency of our queries was very important in enabling us to examine the large cohorts and our large list of features. Finally, we also learned about the importance of a good evaluation metric. We initially thought that classification accuracy would be sufficient, but we soon learned that the nature of our prediction task led to high classification accuracy regardless of our features. In exploring AUC and F_1 scores and their advantages, we ended up being exposed to a much more telling indicator of accuracy. Through our analysis of the data set and user features, we also gained deep insights into early indicators of whether a user will eventually become an elite user. As a result of our work, we hope to equip a site like Stack Overflow with a method to efficiently forecast the future value of new users on a simple set of features and use the results to further target and nurture the growth of elite users.

7.1 Possible Future Work

The next steps for this task involve possible new features and new algorithms. In particular, we could look into temporal information, which involves recreating the user's history. For instance, features involving the time it takes for the first answer to arrive on a new question or the time it takes for the highest scoring answer to arrive could be important for a user's behavior. To better understand the meaning of an elite user, we could look at other metrics and methods of classifying elite users as well. Also, we could explore other algorithms or models to better understand the network. For instance, we thought about looking at this data with a modification of the PageRank algorithm, possibly defining the nodes as users and the edges as high quality interactions between them and producing a PageRank score for each user. We could also look at betweenness centrality to find the most important and connected users in the question and answer network. To generate discussion on the overall structures of the network, we can form clusters and examine the elite user communities. In particular, these methods can either be represented as new models to discover elite users or be used as features in prediction tasks similar to the tasks presented in this paper.

Overall, nurturing the most productive users is paramount to a site like Stack Overflow, where users run, moderate, and produce most of the content. With the large amount of data Stack Overflow collects, we found that simply using some select features could help the site accurately identify the best users within a month of joining the site. In consequence, beyond the normal social evaluators such as reputation, upvotes, and downvotes, we see that there is large potential for predicative power in the features of a user.

8 References

1. A. Anderson, D. Huttenlocher, J. Kleinberg, J. Leskovec. Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow . *In Proc. KDD*, 2012.
2. A. Pal, S. Change, J. A. Konstan. Evolution of Experts in Question Answering Communities. *AAAI*, 2012.
3. X. Liu, W. B. Croft, M. Koll. Finding Experts in Community-Based Question-Answering Services. *CIKM*, 2005.
4. P. Jurczyk, E. Agichtein. Discovering Authorities in Question Answer Communities by Using Link Analysis. *CIKM*, 2007.
5. C. Shah, J. Pomerantz. Evaluating and predicting answer quality in community QA. *SIGIR*, 2010.
6. L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and Yahoo Answers: everyone knows something. *WWW*, 2008.
7. B. Keegan, and D. Gergle. Egalitarians at the gate: One-sided gatekeeping practices in social media. *In Proc. CSCW*, 2010, ACM Press, 131-134.
8. Monard, M. C., and G. E. A. P. A. Batista. 2002. Learning with Skewed Class Distributions. In Advances in Logic, Artificial Intelligence and Robotics, Terceiro Congresso de Lógica Aplicada à Tecnologia, São Paulo - SP, 2002, p. 173-180, IOS Press.