Paper Replication: Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow

Aaditya Bhatia, Abdullah Ahmad Zarir, Daniel Lee, Kundi Yao, Md Hasan Ibrahim
Software Analysis and Intelligence Lab (SAIL)
Queen's University
Kingston, Canada
{aaditya.bhatia, a.zarir, 18dil, 18ky10, ibrahim.mdhasan}@queensu.ca

***Abstract*—Question answering (Q&A) websites offer a plethora of meaningful knowledge that is ready to be unraveled.**

Categories and Subject Descriptors: H.3.4 [**Information Storage and Retrieval**]: Systems and Software.
General Terms: Experimentation, Human Factors.
Keywords: Question-answering, reputation, value prediction.

I. Introduction

Over the last decade, Q&A communities have evolved into a large repository of community-driven knowledge. Notably, communities such as Quora and Stack Overflow have evolved into active and mature communities. In our study, we focus on the Stack Overflow community as it contains one of the most active Q&A communities for developers. The questioner is the user who posed the question. On Stack Overflow, there is a significant fraction of domain experts who can provide answers to questioners with long-lasting value. Since Stack Overflow is a web-based Q&A community, the interactions are stored, so that they can be viewed at any time, which means that the content increasingly has lasting value for users.

Seeing that there is opportunity for long-lasting value for consumers and producers from Q&A communities, techniques can be used to analyze and extrapolate useful information about the community dynamics. Consumers of information are users who utilize the Q&A community to consume knowledge. Producers of information are the domain experts that provide answers to difficult questions on the site. We can guide consumers of information to questions with the potential of having long-lasting value. In addition, we can help producers of information potentially identify questions that have not been successfully answered yet.

Prior works have focused on using question-answer pairs for their analysis. In addition, prior work have proposed approaches to retrieve high quality question-answer pairs with the goal of helping people who have similar questions [].

A holistic view of question-answering sites. Rather than the question-answer pair approach, we alternatively extract information from the community activity by considering questions together with their corresponding set of answers. We view community activity from two levels:

1) **Question level:** We focus on community activity from a question level by using questions with their corresponding set of answer because individual questions have the potential to generate multiple high quality answers. For example, a question such as,"How do you add a remote repository using Git?", can produce multiple high quality replies. We conjecture that questions combined with all their corresponding set of answers can create long-lasting value for questions on sites such as Stack Overflow.
2) **Full site level:** We focus on community activity from a full site level by using the reputation feature from Stack Overflow because reputation provides holistic information about: 1) Levels of community involvement. 2) Incentives for successful contributions and positive behavior. Community involvement and reputation can show us the dynamics of how users provides answers to new questions and how the community approves or disapproves the answer.

Overview of Results. To investigate the potential applications of studying community dynamics from a question and full site level for users on Q&A sites we develop two tasks. The first task is to *predict the long-lasting value* in order to help guide consumers of information to questions with the potential of having long-lasting value. We predict the long-lasting value by computing the question activity within a small time frame after the question is posed. The second task is to *predict whether a question has been sufficiently answered* in order to help producers of information potentially identify questions that may need their contribution.

We use approaches that are constructed by the data from Stack Overflow to address the two tasks. We first identify latent information from the Stack Overflow community. Stack Overflow questions and answers can receive positive and negative votes from community members, which determines the quality of the answer. In addition, the questioner can accept one of the given answers. These factors contribute to a user's

Group assignment for CISC 880.

reputation score, which we use for our analysis.

We identify two principles that provide an organizing framework and features for our two prediction tasks:

1) **Expertise level:** There is a wide range of expertise level that influences the sequence of contributions to a question, with experts generally responding first. The sequence is comparable to a *reputation pyramid*, where experts or elites are at the top of the pyramid and the question trickles down in a top-down manner.
2) **Higher activity level:** Questions with higher activity level signifies the potential interest in the question and the potential of benefitting all answerers based on the evaluation of their answer from the community and their reputation increase. Higher activity questions associate with multiple answerers and can hint at the type of lasting value.

For predicting whether a question will have long-lasting value, we use features based on the answer arrival dynamics within an hour after the question is posed. In doing so, we can classify whether the questions pageviews will be high or low, one year later. We find that number of answers, sum of answer scores, number of questioner's questions, length of highest-scoring answer to arrive, number of comments on highest-scoring answer, and number of comments on highest-reputation answerer's answer are the most powerful features, which shows that attracting a diverse set of answers obtain greater value on Stack Overflow.

For identifying questions that have not sufficiently been answered yet, we predict the questions that offer bounty for a better answer because when a question is not sufficiently answered they will resort to offering bounty. In result, we find that powerful features can lead to an effective prediction.

The main goal of our paper is to use Stack Overflow to provide insights about question-answering sites by leveraging the performance of the features from the two prediction tasks to suggest that community dynamics can provide more information than simple question-answer pairs.

II. Background

A. The Stack Overflow Community

Stack Overflow is one of the most active and successful Q&A sites, where over 90% of the questions receive a response that is accepted by the questioner. More than 80 Q&A sites were influenced and have adopted the same Q&A paradigm as Stack Overflow. In addition, Stack Overflow exhibits qualities that exist in the other Q&A sites: 1) Complex questions on a certain domain. 2) An active community. 3) Significant number of experts.

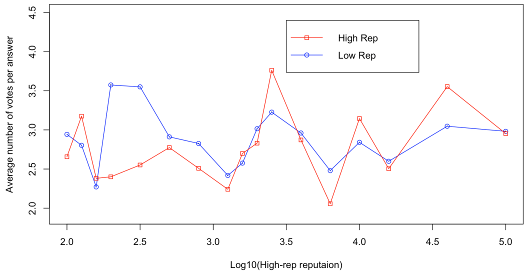

Fig. 1. Average number of votes per answer for both answerers on a 2-answer question as a function of the higher answerer reputation. Lower reputation fixed between 75-125. High reputation plotted on a logarithmic (base 10) scale. (From 2008/07/31 to 2010/12/31)

III. Related Work

IV. Dataset Description

V. Description of tasks

A. Predicting long-term value of a question

B. Predicting whether a question has been sufficiently answered

VI. Community dynamics of question answering

A. A reputation pyramid

B. The activity level of a question

In the previous section, we illustrate the structure of users' reputation pyramid, which explains how answers arrival process relates to the user's reputation, as well as other phenomena from our observation. In this section, we would investigate more on voting, another significant mechanism of Stack Overflow. Except for merely answerers and questioners, other users can also involve in those QA sections by commenting, voting, etc.. The voting mechanism is not only applicable to answers but also to questions, and it is also considered as an important factor to reflect community involvement and evaluation. From our observation of Stack Overflow, we notice that questions with more answers are more likely to benefit from community involvement: answers will get higher votes and questions will receive more favorites. Highly active QA processes are more inclined to benefit from the community, instead of a competence among answerers themselves. Based on our observation, our goal is to validate whether if the activity level is able to explain the feedback and evaluation from the community activities.

Higher activity produces benefits. Like we discussed, unlike the features of general QA sites, Stack Overflow is programming-oriented, the questions are generally hard to be answered by majority community users. Furthermore, based on our observation from the previous section, answerers' arrival time is related to acceptance rate and answerers' reputation, it motivates answerers to answer a question quickly once it is posted. However, is there a competitive relationship existing among answerers?

In order to answer the question above, we choose all the questions with exactly two answers as our target dataset.

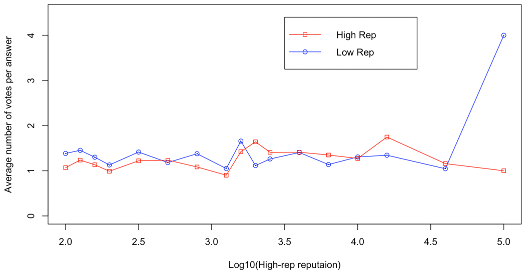

Fig. 2. Average number of votes per answer for both answerers on a 2-answer question as a function of the higher answerer reputation. Lower reputation fixed between 75-125. High reputation plotted on a logarithmic (base 10) scale. (From 2015/07/31 to 2017/12/31)

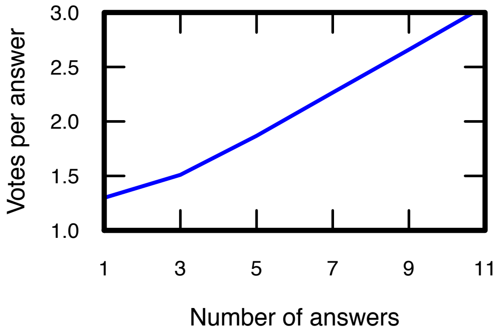

Fig. 3. Number of votes per answer as a function of the number of answers on the question. (From 2008/07/31 to 2010/12/31)

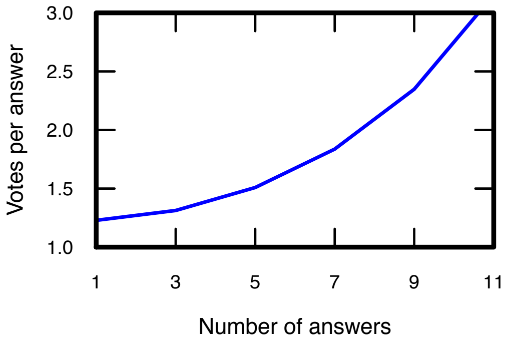

Fig. 4. Number of votes per answer as a function of the number of answers on the question. (From 2015/07/31 to 2017/12/31)

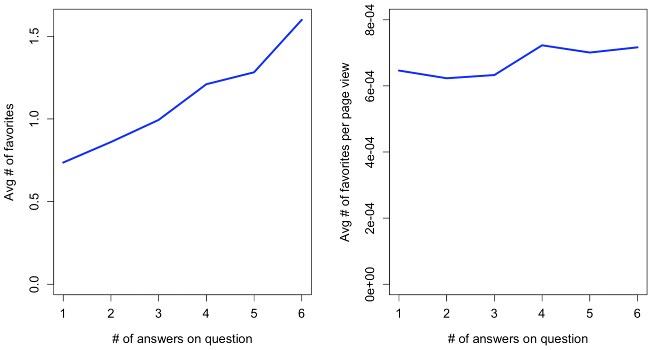

Fig. 5. (Left) Number of question favorites plotted against the number of answers over all questions where the maxi- mum answer vote score is 5. (Right) Same plot normalized by pageviews. (From 2008/07/31 to 2010/12/31)

Suppose r_i is the reputation of the i-th answerer, and v_i is the number of vote score of the i-th answer. If there exists a trend where v_1 goes up while v_2 decreases, we consider there's a competitive relationship between two answerers. Now we set the value of r_i unchanged as the x-axis, and compare the average vote score for both answerers. To begin with, we collected data from all questions with two answers, and separate answers according to answerers' reputation. Since the dataset is highly biased, majority users have very limited reputation, which means they are either non-active users or possibly non-questioner or answerer. Adopting data from this part of users can deviate our result, we set a threshold where lower reputation users of the two answerers should have reputation between 75 and 125. In order to make the graph easier to interpret, first we scale x-axis, which represents higher reputation users' reputation, using a logarithmic(base 10). For each reputation score, we use an average value to present higher and lower answerers' reputation respectively. Instead of representing vote score for each reputation point, we smooth the curve by splitting the vote scores into groups, according to different reputations in the x-axis, and calculate the average value of the group. From Fig1 we notice that in most cases, answer vote scores from high and low reputation users have a similar trend, they either decrease or increase together, but high reputation answerers' answer votes have higher variance. The corresponding pattern reveals that in most cases, the two answerers do not have a competitive relationship. Similarly, this trend can also be noticed during 2015 to 2017. However, we notice the later dataset have relatively less average vote score, this could probably indicate that although there are a great amount of newly introduced answers, the quality of such answers are not as good as before, or the question has been sufficiently answered, newly added answers does not contribute much to the question.

In the following step, we aim to investigate the relationship between the number of answers in each question, to the number of votes per answer. This work is to testify the correctness of the competition theory. If there exists a competitive relationship among answers, we would expect a decreasing number of votes per answer, as the number of answers increases. From Fig 3 we find that from 2008 to 2010, the average number of votes display an ascending trend as the number of answers increases. This fact shows that the hotspot questions (questions with more answers) are more likely to benefit from community attention. Correspondingly, we also notice in Fig 4, where data is collected from 2015 to 2017, we find a similar upward trend for votes per answer as a function of the number of answers. From both of the two datasets we notice that, instead of competing for votes, all answers get more profit from communities involvement, especially for hotspot questions.

Apart from voting volcanism as an important source for

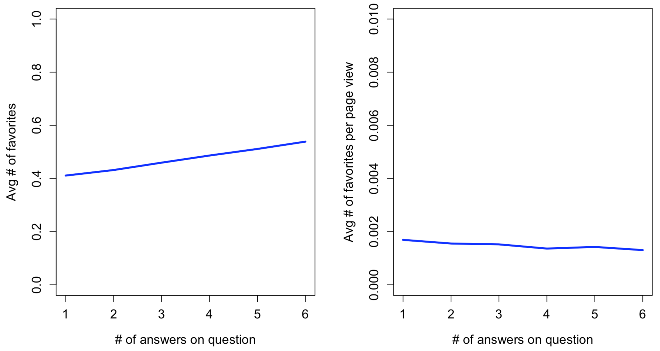

Fig. 6. (Left) Number of question favorites plotted against the number of answers over all questions where the maxi- mum answer vote score is 5. (Right) Same plot normalized by pageviews. (From 2015/07/31 to 2017/12/31)

evaluating community attention, favoring is also another channel to assess the quality of a question. In this part, our concern is whether questions with more answers receive more favorites from the community. In other words, we expect a hotspot question inclines to show its value by gaining more favorites. From previous study [1], Anderson et al. found that among the highest voted answers of all questions, the highest voting score usually receives five votes. In this case, we first narrow our scope to the data from 2008 to 2010, to find all questions with the maximum answer vote score exactly equals to 5. From Fig 5 we can find that the average number of favorites to a question is positively correlated to the number of answers in that question. As a question gaining more answers from community activities, it is probably because the same question is also shared by a wide range of groups. If there is an existing question proposed a commonly possessed confusion, answerers will get more votes for sharing their own experience and solution, and the question will be favored if users are satisfied to acquire instant solutions.

Although we illustrate hotspot questions are more likely to gain a higher number of favorites, the favoriting process usually takes a relatively long time, after a question is posted and even sufficiently answered. During this time period, the potential answerers or users have same concerns will visit those QA pages, this will lead to a significant number accumulation in each question's favorite number. In order to avoid the interference from the number of page views, we normalize the data by dividing the number of favorites from the previous step by the number of page views. From Fig 5 we find the average number of favorites per page view number is quite stable, which means users do not have a significant favoring preference on hotspot questions.

Then we choose questions proposed between 2015 and 2017, where all the data are collected right after the end of 2017, this should reduce the error caused by analyzing questions with more than two years' favorites accumulation. From Fig 6 we notice, even though the trend is similar as we illustrate before. However, we notice a significant scale change between those two figures. From 2015 to 2017, compared with the earlier dataset, we notice a significant decrease in the average number of favorites for all the questions. What's more, we also find the average number of favorites per page view has increased in these years. From our perspective, since we only study the questions posted in recent two years, those questions have relative low favorites number and page views. We think it is because Stack Overflow is already a mature knowledge sharing community after around ten years of development, most of the commonly possessed questions have been sufficiently answered. The newly introduced questions are either duplicating with previous questions or orienting a specific context and requiring expertise in that distinct domain. The duplicated questions will be marked and merged with previously proposed questions, and expertise questions are not accessible to the majority of community users cause they are targeting a special usage scenario.

In this section, we can conclude that answers are not competing with others, all the answers will profit from continuous community involvement. Also, for hotspot questions with many answers, the more answer it gets, the more possible that both questions and their corresponding answers will receive more benefits.

VII. Prediction tasks

A. *Predicting long-lasting value*

B. *Predicting whether the question has been sufficiently answered*

VIII. Conclusion

References

[1] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Discovering value from community activity on focused question answering sites: a case study of stack overflow," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 850–858.