Answering Questions about Unanswered Questions of Stack Overflow

Muhammad Asaduzzaman   Ahmed Shah Mashiyat†   Chanchal K. Roy   Kevin A. Schneider
Department of Computer Science, University of Saskatchewan, Canada
†Department of Computer Science, University of Toronto, Canada
md.asad@usask.ca, mashiyat@cs.toronto.edu, {chanchal.roy, kevin.schneider}@usask.ca

***Abstract*—Community-based question answering services accumulate large volumes of knowledge through the voluntary services of people across the globe. Stack Overflow is an example of such a service that targets developers and software engineers. In general, questions in Stack Overflow are answered in a very short time. However, we found that the number of unanswered questions has increased significantly in the past two years. Understanding why questions remain unanswered can help information seekers improve the quality of their questions, increase their chances of getting answers, and better decide when to use Stack Overflow services. In this paper, we mine data on unanswered questions from Stack Overflow. We then conduct a qualitative study to categorize unanswered questions, which reveals characteristics that would be difficult to find otherwise. Finally, we conduct an experiment to determine whether we can predict how long a question will remain unanswered in Stack Overflow.**

***Index Terms*—Stack Overflow; question-answer; prediction;**

I. Introduction

Unlike a number of community-based question answering sites, such as Yahoo! Answers and Answerbag, which tend to cover a wide range of topics, Stack Overflow (SO) explicitly targets programmers and software engineers. With its inception in 2008, SO has emerged as one of the largest question-answer sites, where community members answer or participate in discussions in pursuit of a solution to a particular problem. Game elements such as *reputation* and *badges* are used to motivate members into participating in site activities. The reputation of a member indicates their site participation in terms of Q&A quality, communication skill, and level of trustworthiness. SO grants highly-reputed community members special privileges to better control site activity. Both questions and answers can be voted on (e.g., up vote, down vote), which serves as a rough measure of Q&A quality. Moderation is not only done by site moderators, but also by members of the community through editing and voting. One important factor behind SO's popularity is its response time [6]. In our dataset we found that 50% of the questions received responses within 15 minutes.

Jeff Atwood and Joel Spolsky started SO, bringing active community members from their own programming blogs (Coding Horror[1] and Joel on Software[2]), which ensured there was a critical mass for the site to be successful. In SO, community members earn privileges when they improve their *reputation*, and can even become moderators alongside the official moderators to ensure a healthy sharing of knowledge on the site. SO's active community attracts information seekers from around the globe harvesting its knowledge-base. Many visitors become members of the site and contribute to the site dynamics.

Although after its first year SO had only around 53K members, the membership increased rapidly and reached close to 1.3 million by 2012. The sharp increase in members and unique tag-base suggests rapid growth for the site. For the first few years SO managed to keep the number of unanswered questions to a minimum, however, site growth has caused an explosion in that number. Figure I shows the number of questions that remain unanswered each month (we considered SO data from July 31, 2008 to July 31, 2012). An unanswered question is a question that is without an answer for at least a month. The figure depicts a rapid growth in unanswered questions over the last two years. To manage the growth, it may be important for SO to introduce a mechanism that helps lower the level of unanswered questions. We seek to understand the characteristics of unanswered questions in order to help moderators better control the site, and to help members posting questions to increase the likelihood of getting answers.

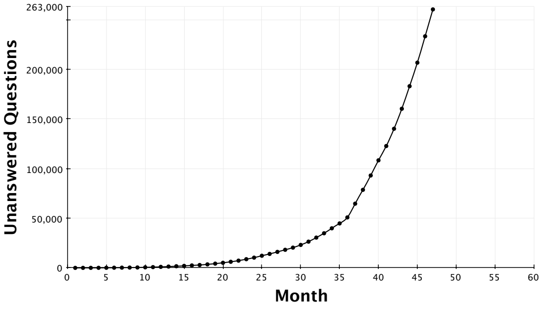

Fig. 1. Number of unanswered questions after each month

In this paper, we mine Stack Overflow data to uncover information about unanswered questions. More specifically, we address the following two research questions:

1) What factors contribute to unanswered questions?
2) Can we predict how long a question will remain unanswered?

[1] http://www.codinghorror.com/blog/

[2] http://www.joelonsoftware.com/

978-1-4673-2936-1/13 © 2013 IEEE

TABLE I
EFFECT OF REPUTATION ON QUESTIONS

Reputation	Users	Asked	Unanswered	Answered
1-100	1159611	1118858	186407	655813
101-1000	105929	1123037	81793	1341438
1001-10000	27414	794760	29510	2707320
10001-100000	2598	114323	2225	1819529
> 100000	68	2812	17	334033

TABLE II
DISTRIBUTION OF ANSWERED AND UNANSWERED QUESTIONS BY DAY OF THE WEEK

Category	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Unanswered	53K	58K	57K	57K	50K	25K	24K
Answered	497K	546K	554K	546K	488K	261K	260K

Towards this goal, we first consider member reputation to see its association with unanswered questions and also present some statistics regarding unanswered questions. Next we manually investigate a random sample of unanswered questions. We use both question content and comments to understand why the questions were unanswered. Finally, we consider how this information can be used to better support the SO site and build classifiers for predicting the number of days a question will remain unanswered. This information can help moderators select questions to edit or route to an expert and can help determine when to offer a bounty.

II. UNANSWERED QUESTIONS IN SO

For the mining challenge [1], we investigated SO data from July 31, 2008 to July 31, 2012. Within this time frame, over 3.4 million questions were asked. While 6.8 million answers were provided for these questions, 299K remain unanswered. We primarily focus on the characteristics of unanswered questions. We consider unanswered questions to be those that remain unanswered for at least one month, resulting in 260K unanswered questions.

Reputation plays an important role in the SO community. From Table I we see that the number of unanswered questions decreases with an increase in reputation. Members with reputation less than 1000 have over 125K unanswered questions while members with more than a 1000 reputation have only a little over 30K unanswered questions. While it is not surprising that novice users ask more questions, it is notable that members with reputation ranges from 1001 to 10000 provide the most answers in SO.

Table II shows a distribution of answered and unanswered questions for the days of a week based on their creation date. Both question categories are significantly less on weekends, maintaining a balanced ratio between them.

III. CLASSIFYING UNANSWERED QUESTIONS

We randomly sampled 400 unanswered questions, selecting 100 questions from each of the four years for which we have data[3]. Commenting on unanswered questions is an indication that users are attempting to answer the question or are providing feedback. It is expected that all questions may not be answered in a crowdsourced Q&A site. However, unanswered questions in SO are still being viewed 139 times on average and this motivated us to investigate why such questions were remaining answered. Our qualitative analysis identified the following characteristics of unanswered questions, which we use as a taxonomy for classifying unanswered questions.

Too short, unclear, vague or hard to follow. One of the main reasons a question remains unanswered is that its meaning is not clear to the community members. Missing information and question format also contribute to not getting an answer. As well, sometimes the code provided with the question is hard to follow. The SO site specifically asks users to be precise when posting a question. However, our analysis shows that a large number of information seekers do not follow this guideline.

Program specific without a program snippet or proper explanation. Often a member asks a question that is specific to a particular scenario in their program. Without seeing the code segment or any detailed explanation of that code it is very hard for another community member to answer the question. For example, in the following scenario, a community member is not provided the appropriate context.

> **Question:** GetPrivateProfileSectionNames always returns 0, even when the file is a valid ini file. What can be the reason?
> **Comment:** 1. Can you provide us with some code? 2. I think the question is too general. What are the values of the parameters you're sending to the function?

Too hard, too specific or too time consuming. We found a few cases where detailed program information was provided, but there was no answer due to the program complexity, the time requirement or even because the question was so specialized that experts on that topic are rare in the community. We found some questions where the members asked information seekers to post the question to an expert mailing list specific to that technology.

Proprietary technology. Questions related to proprietary technology that is not widely used get less attention in community driven web sites, and SO is no exception. Members often redirect them for specific software support.

Impatient, irregular or inconsiderate members. In SO, response turnaround is very fast; members often expect answers within a limited time frame. Impatient users start a conversation but leave after a while making their questions unattended. Sometimes questions and comments are harsh which drives away members from participating in the conversation. For example, there is no post after an asker responds to a comment in the following way.

> **Comment:** Why do you want to read the response twice? This is generally considered bad use, as the code that sends/receives data should be abstracted away, such that consumers just get the data... at that point, you have a reference to the data.
> **Asker:** I am not talking about good or bad practices, I have a clear question about a missing method!

[3] http://tinyurl.com/ak5qouk

Not a question for SO! Community members sometimes share their views or programs in SO and ask for some feedback. These are not meant to be answered, but rather are for getting comments and initiating discussion. Members also post non-programming questions such as questions on management policy, server configuration, system administration, and even hardware and career questions. For example,

> **Question:** I am in my MCA last year. And wished to join educational sector such as lecturer. Simultaneously i need the flavor of industry. So what is the best option available for a person like me who is a part of educational sector but also wishes to work as technical guy.

Often these questioners are asked to post to different forums, most notably superuser.com, which is more general than SO.

Answering their own question. Developers may post their questions as soon as they encounter an issue and participate in a discussion. Often they also post the solution as soon as they find it and leave the question marked as unanswered.

Does not have any answer. Some questions are not answered just because there is no answer. Often this is due to a bug in the software that is revealed after some discussion.

A duplicate question. Community members do not like to see a question which has been posted before. Instead, they would like to see the orginal post updated. They would like an information seeker to do their part reviewing the site before posting a question. SO, however, asked the community not to be harsh to novice members for posting duplicates[4].

Answer no longer relevant or needed. Often members decide to choose a different and even better route to solve the problem they were experiencing. Sometimes they provide some outline of their new solution and sometimes they do not, still leaving the question marked as unanswered.

Fails to attract an expert member. It is not that unusual that a question is being asked properly and still does not get an answer. One of the main reasons for these questions not being answered is incorrect *tagging*. We found quite a few questions with more than five comments without a solution. This characteristic is of concern, since the question is being attended by some members but they are unable to determine a solution. A challenge is how to motivate expert members to respond to an unanswered question.

Works for me! Sometimes an information seeker posts a question with a code fragment that is not working for them. Often the code works for other community members and they are unable to find a problem with the code segment. Often this is due to the asker's system settings.

Course project or homework question. Some members ask for programing solutions that appear to be part of a course project or an assignment. If the question is suspicious, members usually do not answer, abiding by SO rules.

We analyzed the distribution of the 400 randomly selected unanswered questions and categorized each of them. Although, a number of them may fall into different categories, we determine a single category for each question. Table III lists the top five characteristics from the randomly sampled unanswered questions. A number of the questions considered unanswered had answers within their comments; some even with a solution provided as an edit to the post. Nevertheless, it requires more effort to find a solution by scanning comments rather than by directly looking at the solution marked as answered. Since there is no way to close a question or mark a question as answered, a number of the questions considered unanswered actually have answers. Automatically identifying potentially answered questions, which are not marked as answered by a community member, may be a useful feature for SO.

TABLE III
TOP FIVE CHARACTERISTICS OF UNANSWERED QUESTIONS

Percent	Characteristic
21.75	Fails to attract an expert member
17.0	Too short, unclear, vague or hard to follow
12.0	A duplicate question
11.75	Impatient, irregular or inconsiderate members
9.0	Too hard, too specific or too time consuming

IV. How long will a question remain unanswered?

We hypothesize that the reputation of a question asker, their association with other community members, and votes received for a question and accompanying code are a number of factors that contribute to the chance of getting an answer. Unlike Yahoo! Answers where questions have a four day open period, SO does not have such a restriction. Unanswered questions in SO can be answered any time in the future. Thus, SO moderators or information seekers will benefit more if we can predict how long a question will remain unanswered. Moderators can then take further action on those questions that are expected to have long response times. A response time estimate can also help askers decide whether they should offer a bounty when they post their question in order to encourage more visitors. Since the majority of questions (89.6%) in SO received their first answers within a day, we focus our attention on questions that remain unanswered for more than a day.

In order to build prediction models we count only those questions in SO that receive at least one answer and that remain unanswered for more than one day. We count the number of different votes (e.g., up vote, down vote, favourite vote) received by each post within the first day of its posting, since votes are a rough measure of a question's popularity and popular questions are likely to have shorter response times. We also consider the following heuristic measures for each post.

1) **Title Length:** Number of characters in a post title.
2) **Post Length:** The line length of a post (after removing the code block and HTML tags).
3) **Tag Similarity:** The number of previous posts with tags similar to this post. In order to reduce the effect of highly popular tags we consider a minimum value among all the tags used for a post.
4) **Readability:** We hypothesize that questions that are difficult to read and understand will remain unanswered

[4] http://meta.stackoverflow.com/questions/9953/could-we-please-be-a-bit-nicer-to-new-users

TABLE IV
PREDICTION MODEL

Random Forest					
Precision			Recall		
C_1	C_2	C_3	C_1	C_2	C_3
0.351	0.344	0.374	0.403	0.353	0.329

J48					
Precision			Recall		
0.361	0.349	0.378	0.351	0.295	0.447

for a longer period of time (or take more time to be answered). We consider four basic readability measures (i.e., ARI, Coleman, Flesch, and Fog) in this study based on McCallum and Peterson [2].

5) **Code:** A nominal attribute indicating whether a post contains a code snippet or not.

We consider data about askers for each post:

6) **Question Answered:** Number of questions answered by an asker before posting a question.
7) **Question Asked:** Number of questions asked by an asker before making a post.
8) **Asker Score:** The difference in number of up votes and down votes received by the asker before the post. Votes are collected for both the questions and answers posted by the asker.
9) **Satisfaction:** The number of previous answers provided by the asker and accepted by other members.

We use an equal-frequency binning scheme to categorize the time questions remain unanswered (in minutes) into three categories: C_1: $1440m<t<=5570m$; C_2: $5570m<t<=39285m$; and, C_3: $t>39285m$ $(27.28\ days)$. We did not categorize beyond 27.28 days response time, since we believe that such a response time is undesirable for an asker. We built prediction models with the Weka toolkit[5]. To train the classifiers we used the last two years of SO data and we used 10-fold cross-validation to evaluate the prediction models. Table IV shows the precision and recall measures for two different classifiers. Although our results requires further improvement for practical use and needs more experimentation, it shows promise that we can predict response time using attributes of SO posts.

V. Related Work

Prior work on community Q&A services focused on studying the dynamics of community activities. Anderson et al. [4] used such knowledge to predict the long lasting value of a question and to decide whether a question has been sufficiently answered. Admaic et al. [3] leveraged community knowledge to predict the best answer. Harper et al. [5] compared answer quality of community-based question answering sites and found that topic, type, and payment are the significant contributors for getting the best answers. Yang et al. [9] applied a machine learning technique to predict not-answered questions in Yahoo! Answers. However, the policy that governs not-answered questions is not the same for Yahoo! Answers and SO. Whereas Yahoo! Answers has a 4 day time limit for answering questions, there is no such restriction in SO.

Mamykina et al. [6] studied interaction and usage patterns in SO followed by a qualitative study to understand factors contributing to its success. They found that tight engagement of official moderators, offering competitiveness through a reputation mechanism, and supporting continuous feedback, collectively helped SO become a successful information hub. Treude et al. [8] presented a classification of questions asked in SO. They identified that *how-to* and *conceptual questions* are the most popular ones.

While previous work focused on answers or site dynamics, we focused on unanswered questions. The closest research to our study is the work of Nasehi et al. [7], who studied SO Q&A threads to understand characteristics of good code examples. They identified nine attributes of recognized answers and also pointed to three factors that contributed to low-voting. Our study differs from theirs in that we want to understand why questions are being unanswered in SO.

VI. Conclusion

7.5% of SO's total questions are unanswered. We provide a taxonomy for this minority group to help understand why there is a delay in answering questions and whether a question requires further information. Using the taxonomy, post metrics and information about askers, we built a classifier that predicts how long a question will remain unanswered in SO. Moderators can use the classifier to predict questions likely to have delayed responses and can consider taking actions to promote an earlier response, such as editing the question or routing the question to an expert member. Identifying the quality of a question can also help in this regard, which we plan to consider in a future study.

References

[1] A. Bacchelli, "Mining challenge 2013: Stack Overflow", in Proc. MSR, to appear, 2013.

[2] D. R. McCallum, and J. L. Peterson, "Computer-based readability indexes", in Proc. ACM, pages 44–48, 1982.

[3] L. Adamic, J. Zhang, E. Bakshy, and M. Ackerman, "Knowledge sharing and yahoo answers: everyone knows something", in Proc. WWW, pages 665-674, 2008.

[4] 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 Proc. KDD, pages 850–858, 2012.

[5] F. M. Harper, D. Raban, S. Rafaeli, and J.A. Konstan, "Predictors of answer quality in online Q&A sites", in Proc. CHI, pages 865–874, 2008.

[6] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann, "Design lessons from the fastest Q&A site in the west", in Proc. CHI, pages 2857–2866, 2011.

[7] S. M. Nasehi, J. Sillito, F. Maurer, and C. Burns, "What makes a good code example? A study of programming Q&A in StackOverflow", in Proc. ICSM, pages 25–35, 2012.

[8] C. Treude, O. Barzilay, and M. Storey, "How do programmers ask and answer questions on the web?", in Proc. ICSE, pages 804–807, 2011.

[9] L. Yang, S. Bao, Q. Lin, X. Wu, D. Han, Z. Su, and Y. Yu, "Analyzing and predicting not-answered questions in community-based question answering services", in Proc. AAAI, pages 1273–1278, 2011.

[5] http://www.cs.waikato.ac.nz/ml/weka/