Aren't Those Questions Interesting Enough?
Investigating the Root Causes of Unanswered Questions

Ripon K. Saha* Avigit K. Saha† Dewayne E. Perry*
* The University of Texas at Austin
ripon@utexas.edu, perry@mail.utexas.edu
†University of Saskatchewan
avigit.saha@usask.ca

***Abstract*—Stack Overflow is a highly successful question-answering website in the programming community, which not only provide quick solutions to programmers' questions but also is considered as a large repository of valuable software engineering knowledge. However, despite having a very engaged and active user community, Stack Overflow currently has more than 300K unanswered questions. In this paper, we mine Stack Overflow dataset to investigate why these questions remain unanswered by applying a combination of statistical and data mining techniques. Our results indicate that although there are some topics that were never answered, most questions remained unanswered because they apparently are of little interest to the user community.**

I. Introduction

Recently, the community question-answering site has become a popular media for information exchange. These sites leverage the knowledge and expertise of users to provide answers, whose end product has a long lasting value. `Stack Overflow` is such a leading question-answering site in programming community, where developers can ask and answer programming related questions. As of July 2012, `Stack Overflow` had 3.45 million questions with a mean arrival rate of 5.6K questions per day. Among them, more than 90% questions have at least one answer within a median time of 12 minutes [1]. However, while the proportion of unanswered questions is small (approximately 10%), that still leaves a substantial number of unanswered questions (approximately 300K questions). Furthermore, the proportion of unanswered questions is increasing every year (see Table I).

Programmers generally post questions to `Stack Overflow` when they are stuck on some points and have possibly no coworkers to help. The hope is that they will get a quick solution or suggestion from some fellow expert in the same community for the given problem. Therefore, it can be very frustrating and impede their normal development progress if they do not get an answer for their question. Given that all questions are meant to be objective and factually answerable in `Stack Overflow`, we believe it is important to investigate why such a large volume of questions remains unanswered.

There may be, among others, the following three reasons that a question remains unanswered. First, it may be that the experts who are willing and able to answer the question have not noticed it. This may occur for a variety or reasons– for example, inappropriate tagging, or the questions were posted in the weekend, etc. Second, the question may not be interesting enough and the potential answerers have just ignored it. Third, the question may be very hard and no one in the community knows an appropriate answer.

Investigating a large volume of data presents a significant challenge because a meaningful manual investigation is literally impossible. In this paper, we apply a combination of statistical and data mining techniques to systematically investigate the possible reasons of a question being unanswered. To this end, we first encode each question with a set of attributes, which we call a *feature vector*. We then delineate which attributes are more important than the others in differentiating answered questions (AQ) from unanswered questions (UQ) and test whether we can use those attributes to predict *UQ*s in advance. Finally, we investigate the topics that have not been answered yet to get an overview about how frequently such topics occur. To the best of our knowledge, our study is the first to characterize the unanswered questions systematically and comprehensively.

The rest of the paper is organized as follows. Section II describes the `Stack Overflow` dataset. In Section III, we describe how we encode each question into a *feature vector*. Section IV presents techniques used to characterize unanswered questions along with results. Finally Section V concludes the paper with our directions for future research.

II. Dataset Description

A user can perform a wide variety of functions on `Stack Overflow`. Among them, the most basic functions are asking and answering questions. Both the questions and answers can be upvoted or downvoted by other users. The difference between these up votes and down votes for a given question/answer are actually used to determine the importance ($score = upvotes - downvotes$) of that question/answer. Furthermore, users can mark questions as their favorites. The

TABLE I
Proportion of Unanswered Questions by Year

Year	AQ	UQ	[%]
2008	61,480	100	0.16
2009	350,310	2,799	0.79
2010	698,386	17,481	2.44
2011	1,176,422	102,207	7.99
2012	866,980	173,413	16.67

TABLE II
GENERAL STATISTICS

Facts	Description
Questions	3,453,742
with an accepted answer	2,148,306
without an accepted answer	1,005,272
not answered	300,164
Answers	6,858,133
Comments	13,252,467
Registered Users	1,295,620
Asked	625,110
Answered	443360
Commented	484828

questioner can also select an answer as the accepted answer, which indicates that it is the best answer for the given question.

`Stack Overflow` also has a reputation system to encourage users to produce high-quality content and to be engaged with the site. Whenever users provide a meaningful answer that is upvoted by other users or accepted by the questioners, they gain some reputation. On the other hand, users can lose their reputation if any provided questions/answers are downvoted or marked as spam. The reputation score of a user actually represents how useful he/she is for the community and determines his/her privileges on the site.

We have used the complete trace of all the aforementioned actions on the `Stack Overflow` website between its inception on July 31, 2008 and July 31, 2012 provided by the mining challenge organizers [2]. The dataset contains descriptions of different posts (e.g. questions, answers, or comments), users, votes, and so on. Table II presents some of the basic statistics of the dataset relevant to our study. However, we excluded all the questions that are: i) *closed*–either for duplications or not useful, or 2) *posted in the last two days* of the database. Since we are investigating why a question remained unanswered, we assumed the questions posted in the last two days may not have gotten enough time for an answer. We have chosen two days as a threshold because `Stack Overflow` does not permit a questioner to spend bounty points before two days have passed. We feel that this is more important than the fact that a question is answered, typically, within a median time of 12 minutes [1].

III. ENCODING QUESTION CHARACTERISTICS

In order to study the characteristics of unanswered questions, we encode each question into a *feature vector*, which consists of a set of attributes. Overall we explore three different classes of attributes. This section introduces each attribute and the rationale of choosing that attribute.

Structural Attributes: We first explore the attributes that are related to the question itself and that may affect the possibility of getting an answer. For example, tags are used to categorize questions so that one can find his/her questions of interest easily. A user can also set a tag-based notification, i.e., whenever a question is posted associated with a tag that he/she is interested in, the user will be notified. Therefore, appropriate tagging of questions may increase the possibility of getting an answer. Similarly, some busy users may be reluctant to answer very long or vague questions. We select four features in this class:

- a_1 : Number of Tags (1 to 5)
- a_2 : Length of Questions
- a_3 : Presence of Code (Yes/No)
- a_4 : External Link (Yes/No)

Quality Attributes: While the aforementioned attributes give us some idea about the structure of the question, there are a rich set of dynamic attributes that can give useful hints about the quality of the question. For example, we can assume that the higher the number of views, scores, and number of favorites of a question, the more important the question is to the community. We select four features in this category.

- a_5 : Number of views
- a_6 : Score
- a_7 : Number of favorites
- a_8 : Number of comments

Questioner Attributes: The history of a user who asked a particular question may provide useful information as to whether a question will be answered. It is highly likely that a person with deep knowledge about some area will ask high quality question. We select four attributes about the questioner.

- a_9 : Reputation
- a_{10} : Number of answered questions in the past
- a_{11} : Number of unanswered questions in the past
- a_{12} : Percentages of questions got answers in the past

IV. CHARACTERISTICS OF UNANSWERED QUESTIONS

This section presents our methodologies and results towards understanding the characteristics of *UQ*.

A. Range, Central Tendency, and Standard Deviation

In order to investigate the characteristics of *AQ* and *UQ*, first we measure the ranges, central tendencies, and standard deviations of the relevant attributes for both *AQ* and *UQ* separately. The results presented in Table III show that the range of each attribute for *AQ* subsumes that of *UQ*. This is expected because of the large proportions of *AQ* compared to *UQ*. However, we observe that the mean value of most quality attributes (views, score, favorites) and questioner's reputation for *AQ* are clearly greater than that of *UQ*, which indicate that *UQ* are relatively less interesting than *AQ*. On the other hand, mean size of *UQ* is greater than that of *AQ*. Since attributes a_1, a_3, and a_4 are nominal data we excluded them from this type of measurement.

B. Frequency Distribution

Although the aforementioned basic statistics provide a very good idea about which attributes are more useful than others in differentiating *AQ* from *UQ*, it is difficult to conclude anything because the data is highly skewed. Therefore, we investigated the frequency distributions of the promising attributes (from the previous section) to understand them in more detail. We intuitively choose an appropriate interval length for each attribute to count the number of questions. It should be noted that the total number of *AQ*s is almost 10 times greater than that of *UQ*s in the dataset, which is equivalent to one vertical scale in the graph. Therefore, if the frequency of *AQ* is only 1 scale higher than that of *UQ* at any given point, the probability

TABLE III
RANGE, CENTRAL TENDENCY, AND STANDARD DEVIATION

Attributes	Answered Questions					Unanswered Questions				
	Min	Max	Mean	Median	Std. Dev.	Min	Max	Mean	Median	Std. Dev.
Length of Questions (a_2)	5	48,258	1,079	711	1,389	19	35,588	1,300	780	1,845
Number of Views (a_5)	1	1,051,784	789	228	3,441	2	58,573	141	83	316
Score (a_6)	-132	2,499	1.62	1	7.02	-14	264	0.27	0	1.03
Number of Favorites (a_7)	0	5,894	2.17	0	13.13	0	20	0.9	0	0.64
Number of Comments (a_8)	0	109	2.72	0	2.36	0	38	2.82	5	2.22
Questioner Reputation (a_9)	1	465,166	1,886	338	7,005	1	223,117	579	46	2,586

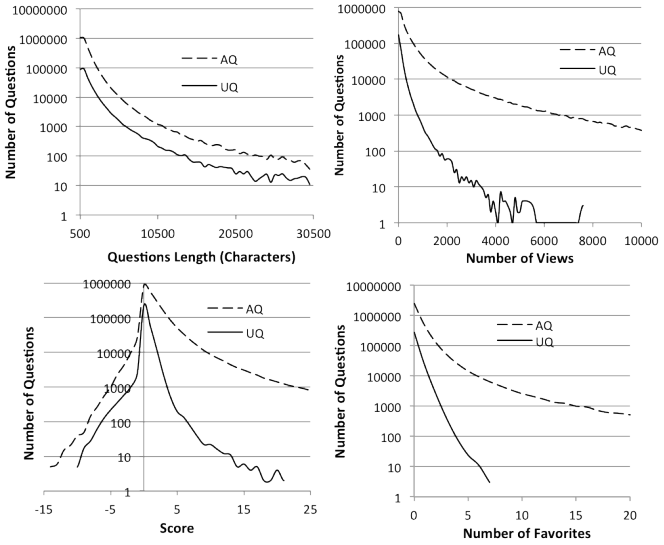

Fig. 1. Frequency Distribution of Question Length, Number of Views, Score, and Number of Favorites

of getting either *AQ* or *UQ* at that point is literally the same. From Figure 1, now it becomes evident that number of views, and number of favorites are clearly greater for *AQ* than *UQ*. The question scores also follow the same trend. Although some questions were answered with negative scores, as the score increases the proportion of *AQ* also increases. In fact, we have found only 89 *UQ*s in total having a score more than 10. These findings indicate that almost all the questions that are interesting to the community get answers. Finally, we observe that although the mean length of *UQ* is reasonably greater than that of *AQ* (from the previous section), the probability of getting an *AQ* and *UQ* at any given length is the same since *AQ* always maintained 1 scale difference from *UQ*. We have also investigated the frequency distributions of a_1, a_3, and a_4 in *AQ* and *UQ* but found no systematic differences.

C. Ranking Features

In the previous section, we showed that there are certain attributes (a_5, a_6, a_7, a_9), whose values are different for *AQ* and *UQ*. This finding indicates that these attributes are key in predicting whether a question will be answered or not. However, we do not know yet which attributes are more important than others in differentiating *AQ* and *UQ*. Therefore, a ranking of these attributes would be very helpful to select the *top n* attributes for the prediction task, where the value of n will be selected by the system according to the required precision level. Reducing the number of these attributes is important because learning the appropriate values from millions of questions in such a high dimensional space is computationally very expensive. We use two popular statistical measures–*information gain* and *information gain ratio*–based ranking algorithms to rank our attributes.

1) Information Gain: In information theory, the information gain of a random variable is the change in information entropy from a prior state to a state that takes the variable as given. Therefore, the information gain of a particular attribute in classifying if a question is *AQ* or *UQ* is:

$$InfoGain(C, a_i) = H(C) - H(C|a_i) \quad (1)$$

where C represents a particular class (*AQ* or *UQ*), a_i denotes the attribute, and H denotes information entropy.

2) Gain Ratio: Although information gain is usually a good measure for deciding the relevance of an attribute, it favors the attributes that can take on a large number of distinct values. Therefore, we have used gain ratio to rank our attributes, which overcomes the previous problem. Gain ratio is mathematical defined as Equation 2, where all the symbols are as previous.

$$GainRatio(C, a_i) = \frac{(H(C) - H(C|a_i))}{H(a_i)} \quad (2)$$

However, gain ratio gives an unfair advantage to the attributes with very low information values. Therefore, we have used both rankings to obtain a balanced result.

3) Data Balancing: A major problem in most of the data mining applications is unbalanced data because machine learning algorithms can be biased towards the majority class due to over-prevalence. In our study, we also observe that our dataset is highly unbalanced. There are 90% of total questions in the answered category, whereas it is only 10% in the unanswered category. Therefore, we first need to balance the dataset.

Oversampling the minor category or undersampling the major category are the two common ways of balancing dataset. However, both methods have some drawbacks. Oversampling introduces a bias towards the minor category, whereas undersampling may exclude useful corner cases. Therefore, we have used a selective sampling method, which is best suited for our approach. In order to selectively sample our dataset, we have considered only those questions in the *AQ* category if (i) there are more than three answers to the question, and (ii) there is an accepted answer. Furthermore, we have also excluded all the questions from both categories (*AQ* and *UQ*) where the questioner user id is not available. This sampling process gives us 329,840 questions in *AQ* (55%) and 272,719 questions in *UQ* (45%) category, which is a fairly balanced dataset. Since the number of answers to a question is not a considered attribute in our study and more answers implies better question quality, this sampling process also give us high quality data for the learning and ranking purposes.

TABLE IV
FEATURE RANKS

Rank	Information Gain		Information Gain Raio	
	Attribute	Score	Attribute	Score
1	Views (a_5)	0.364	(a_6)	0.136
2	Score (a_6)	0.339	(a_7)	0.101
3	User Reputation (a_9)	0.271	(a_5)	0.071
4	Favorites (a_7)	0.142	(a_9)	0.051
5	Percentages (a_{12})	0.109	(a_{12})	0.028
6	Unanswered Questions (a_{11})	0.056	(a_{11})	0.021
7	Question Length (a_2)	0.021	(a_2)	0.006
8	Answered Questions (a_{10})	0.015	(a_4)	0.006
9	Has Code (a_3)	0.003	(a_{10})	0.004
10	Has Link (a_4)	0.003	(a_3)	0.004
11	Comment Count (a_8)	0.002	(a_8)	0.001
12	Tags (a_1)	0.001	(a_1)	0.001

4) Result: We use the Weka [3] implementation of Information Gain Ranking and Gain Ratio Ranking algorithm with default settings to rank the attributes defined in Section III. Table IV presents the detailed ranking results with their corresponding scores. From the both rankings, we see that the number of views, question scores, and questioner reputation are the most dominant attributes in deciding whether a question is *AQ* or *UQ*. Although there are some differences between two rankings, attributes in Top 7 are the same.

D. Predicting Unanswered Questions

To see if we can predict whether a question is *AQ* or *UQ*, we build a prediction model using the top 6 attributes from Table IV. This experiment also justifies the effectiveness of our ranking to reduce irrelevant features. We use the Weka implementations of the C4.5 decision tree [4] learner (known as J48) to build the prediction model. Decision trees are the most widely used machine learning algorithms, and perform well with large data in a short time. They perform a general to specific search of a feature space, adding the most informative features to a tree structure as the search progresses. The objective is to select a minimal set of features that efficiently partitions the feature space into classes of observations. Decision trees are simple to understand and interpret, and can handle both numerical data and categorical data.

We use 10-fold cross validation to evaluate our prediction model based on two metrics, precision and recall. Our results show that our model can predict the *UQ* with a precision of 0.88 and recall of 0.91, both of which are highly accurate. The weighted precision and recall for both *AQ* and *UQ* are 0.9 and 0.89 respectively. We have also built another prediction model using all 12 attributes and got the precision of 0.89 and recall of 0.91 in identifying *UQ*. Therefore, the results show that using the top 6 attributes from the rankings, we only lose precision of 0.01 and lose nothing in recall. Furthermore, the computation time to derive the results using top 6 attributes is significantly less than using all 12.

In order to see if other classifiers can increase the accuracy of our prediction, we have used three other popular classifiers– K-Nearest Neighbor, Naive Bayes, and Random Forest. However, decision trees outperformed all of them in predicting *UQ* in terms of precision while maintaining high recall. The overall F-Measure is also the best for decision tree. Table V shows the detailed result.

TABLE V
ACCURACY OF PREDICTIONS USING THE TOP 6 FROM THE RANKING

Classifiers	AQ		UQ		Overall
	Precision	Recall	Precision	Recall	F-Measure
J48 Decision Tree	0.92	0.89	0.88	0.91	0.90
K-Nearest Neighbour	0.78	0.91	0.87	0.81	0.81
Naive Bayes	0.98	0.56	0.65	0.98	0.74
Random Forest	0.96	0.77	0.78	0.96	0.85

TABLE VI
UNANSWERED TOPICS

Tag Name	Freq.	Tag Name	Freq.
jquery-jtable	8	jmyron	4
lineseries	5	purepdf	4
avplayerlayer	4	glog	3
ace-datatable	4	scroll-paging	3
fxcomposer	4	timeglider	3

E. Unanswered Topics

From the previous sections, we observe that the quality attributes are the most significant factors in deciding whether a question will be answered or not. However, to see if there are any uncommon topics and how often they are asked, we investigated the topics that were never answered. Since, manually investigating the topic of each question is practically impossible, we considered the question tags as the representatives of question topics. Then we searched the distinct tags that are present in *UQ* but not in *AQ*, and counted the number of questions associated with those topics. We have found 274 unanswered topics. However, most of the topics appeared in a single question. As a result, we have found only 378 questions in total associated with those topics, which is almost negligible compared to the large number of unanswered questions. Table VI shows the top 10 unanswered topics with their frequencies.

V. CONCLUSION

`Stack Overflow` has more than 30,000 tags, which cover a diverse variety of topics, from very general to very specific, in the software development domain. Among them, we have found only 274 topics, and 378 questions in total associated with those topics that were not answered. This finding indicates that there is at least an expert for 99% of the topics. Therefore, the possibility of getting such a huge number of unanswered questions for lack of experts is literally very small. We also have not found any noticeable relationships between structural attributes and possibility of getting answers. On the other hand, we have found that the quality attributes, questioners' reputation and previous history are very useful in predicting whether a question will be answered or not. Therefore, it seems that the unanswered questions are of little interest to the user community. We feel, however, that further study is necessary to explore the issues in more depth and to strengthen our conclusions.

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," Proc. *KDD*, 2012, pp. 850–858.

[2] A. Bacchelli, "Mining challenge 2013: Stack overflow," Proc. *MSR*, 2013, to appear.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *ACM SIGKDD Explorations*, 11(1):10–18, 2009.

[4] J. R. Quinlan, "C4.5: Programs for Machine Learning," *Morgan Kaufman*, 1993.