2015 12th Working Conference on Mining Software Repositories

ETA: Estimated Time of Answer Predicting Response Time in Stack Overflow

Jeffrey Goderie,* Brynjolfur Mar Georgsson,† Bastiaan van Graafeiland,‡ Alberto Bacchelli§
{*c.j.m.goderie, †b.m.georgsson, ‡b.vangraafeiland}@student.tudelft.nl, §a.bacchelli@tudelft.nl
Delft University of Technology, The Netherlands

***Abstract*—Question and Answer (Q&A) sites help developers dealing with the increasing complexity of software systems and third-party components by providing a platform for exchanging knowledge about programming topics. A shortcoming of Q&A sites is that they provide no indication on when an answer is to be expected. Such an indication would help, for example, the developers who posed the questions in managing their time. We try to fill this gap by investigating whether and how answering time for a question posed on Stack Overflow, a prominent example of Q&A websites, can be predicted considering its tags. To this aim, we first determine the types of answers to be considered valid answers to the question, after which the answering time was predicted based on similarity of the set of tags. Our results show that the classification is correct in 30%-35% of the cases.**

I. Introduction

With software becoming increasingly influential in our lives, new technologies that trigger new problems and new questions are 'born' at any time. To effectively handle this situation Q&A sites come to the rescue, offering a platform for exchanging knowledge about programming topics. Stack Overflow (SO) is the most prominent example of such sites.

Although previous research showed that most SO questions are answered in a median time of 11 minutes [5], this depends largely on the topic [3] and Q&A website offer no relevant ETA (estimated time to answer) information. Having an indication for ETA would help, for example, developers better schedule their working time (*e.g.*, they could decide whether to switch focus to a different development part while waiting for the answer, or to take a break and wait for the answer). In this paper, we investigate whether we can compute an ETA of a SO post. In particular, we focus on a prediction based on post's tags, as they are the most relevant feature to define the topic of the question and are user defined.

We start our investigation by investigating which answer we have to consider as the valid one, whose time we want to predict. In fact, Bosu *et al.* showed that only considering the first accepted answer as the valid one might prove biased [4] and Asaduzzaman *et al.* reported that one of the reasons why questions remain unanswered on SO is that the asker never selects an accepted answer [2]. Subsequently, we find a tag-based metric to group similar questions and devise how to measure our prediction's effectiveness.

Our results show that we can provide a correct prediction at least 30% of the time, and if we assume that an answering time 6 times smaller or bigger than the predicted one is acceptable, as done by Arunapuram *et al.* [1], we are able to give reasonable predictions over 60% of the time.

II. Related Work

We briefly describe prior attempts at predicting answering times of Q&A posts.

Metrics. Bhat *et al.* investigated post's features in relation to response time [3]. By predicting which questions (1) will receive an answer within the median response time and (2) are answered within an hour, they found that tag-based features show the highest correlation to the response time among the researched features. In our work, we strive for a more fine-grained computation of ETA and we adopt a different approach for filtering answers and posts.

Accepted answers versus community answers. Rechavi *et al.* researched the differences between user accepted answers and community accepted answers in Yahoo! Answers [6]. While this service handles questions slightly differently than SO, they found a characteristic that works for Q&A sites in general: Commonly askers go for satisfying results, while the community goes for the best result.

Unanswered questions. Asaduzzaman *et al.* investigated the unanswered questions of SO [2], finding that 12% of these remain unanswered because askers do not always return to select an accepted answer. Other top reasons why questions remain unanswered that they reported are: failure to attract experts, unclear or duplicate questions, and rude users.

Measuring success. Arunapuram *et al.* analyzed the distribution and the correlation of response times in SO according to various features, like title length and word use, and tried to use these to predict response [1]. They measured their results in relative error (*i.e.*, the relative difference between predicted time and actual time) and by focusing on accepted answers and first answers they obtained relative errors of 0.40 and 1.36 respectively, by using most frequent occurring time range as a predictor. They discovered that using tags led to the lowest relative errors. While they focused on only 100 tags, using a seemingly computational-heavy algorithm, we use more tags and simple algorithms to do predictions.

III. Methodology

In this section we describe the research method that we followed in our investigation on predicting ETA based on tags information from two sources.

978-0-7695-5594-2/15 $31.00 © 2015 IEEE
DOI 10.1109/MSR.2015.52

A. Finding relevant answer types

As first answers can be affected by the ‘fastest gun in the West’ problem [4], we investigate other ways for picking the answer whose time we want to predict.

Analyzing the complete SO dump [7], which contains 8.7 million questions, we find that almost 5 million questions have an accepted answer. Inspired by Rechavi *et al.* [6] on community vs. asker selected answers, we compared accepted answers to the highest scoring ones and found that their scores overlap in 90% of the cases. This makes it reasonable to consider the highest scoring answer as the right one; if multiple answers had the same score, or all scores were 0, the earliest answer with the highest score was considered. By making this assumption, we would have a significantly larger data set to our disposal, with 7.7 million posts. However, it generally takes longer for a community to select the best answer than it takes for the accepted one [6], thus we consider these two types of answers separately for the follow-up steps.

B. Computing tag-based features

To predict ETA based on tags’ characteristics, we get inspiration from the features defined by Bhat *et al.* [3], who found three tag-related features to be significantly related to response time: (1) Active subscribers ratio (ASR) per tag, (2) Responsive subscribers ratio (RSR) per tag, and (3) Popularity rating (PR) per tag.

ASR is the ratio of users who posted “sufficient” answers to questions with the considered tag within the “recent past” over the total amount of users. We consider 10 answers within the recent past (see Section III-C) as it resulted in the smallest number of tags with no active subscribers. RSR is the ratio of users whose average response time for questions with the considered tag is less than an hour (2 times the median answer time) in the recent past over the total amount of users. PR is the ratio of how often a specific tag occurs over the total amount of tags. It is reported to have a strong correlation with answering time [3]. Bhat *et al.* computed PR over all the questions in the whole life of SO, while we only consider tags the recent past (see Section III-C), to better account for time trends in tags.

These features involve different information sources (to respect the requirements of this year’s MSR Challenge): The first two are user-based and the last is tag-based. In our investigation we analyze these features separately and in conjunction, by information source.

A post with n tags has n values for each tag-based feature; to aggregate into a single value we experimented with both sums and averages. Since results were extremely similar, we only report the ones considering the average.

C. Collecting data

To ensure that our results were not largely influenced by trends in tag popularity, we limit our analysis to questions that posted between May 2014 and Aug 2014. As such, our ‘recent past’ is defined as a 3-month period; this ensures that the popularity of the tag examined, as well as the active and responsive subscriber ratios, are up to date. We discarded every question that had an answer count of zero.

From earlier attempts we found a number of posts with sporadically-used tags that had a large variance in answering times, which negatively influence predictions. To remove this influence, we discarded tags with less than 15 unique responders (close to the upper quartile) within our time frame. This also makes us focus only on strongly represented tags and reduced the tags from 25,000 to 6,500.

We computed answering time as the difference between creation dates of questions and user-accepted/highest-scoring answers. As done by Arunapuram *et al.* [1], we divide answering times in 25 time bins using the k-means clustering.

Through the data collection we obtained 3 data sets: One dataset with tag-based metrics for more than 6.5 thousand tags, the ‘COMMUNITY-BASED Answers’ dataset with 444,000 posts, and the ‘USER-ACCEPTED Answers’ dataset with 263,000 posts.

D. Calculating the ETA

To predict the bin of each post and evaluate the outcome, we conduct 10-fold cross validation and use the supervised learning algorithm k-Nearest Neighbors. This algorithm selects k posts from the training set whose metrical values are “closest” to the metrical values of the post from the testing set by using the Euclidean distance. After the k nearest neighbors are found, the post gets assigned to the most frequently occurring bin among them. By investigating different k values, we found that k does not significantly influence performance when selected large enough. We set a final k value of 10, to prevent k-means clustering bias from perpetuating onto the k-NN algorithm.

We considered two measurements of success: (1) rate of successful prediction and (2) error between actual and predicted time. The former determines the ratio of successful predictions over the total amount of predictions (*aka* successful classification rate). In case of a low classification rate, we want more detailed insights on how the prediction worked, for this we use the latter metric, defined by Arunapuram *et al.* [1] as:

$$\frac{|Actual\ time - Predicted\ time|}{\min(Actual\ time, Predicted\ time)}$$

The actual time is known exactly but the predicted one is a range, thus we consider as predicted time the median value of the predicted bin, which is the most representative of the bin.

IV. RESULTS

Our aim in this section is to evaluate to what extent we succeeded to accurately predict answering times on Stack Overflow. In general, a prediction is assumed to be accurate if the predicted answering time is in the same bin as the actual answering time. In section 3 we explained how we filtered the data to obtain 3 data sets. We also defined several metrics to express how successful the approach was. The presented results will cover each data set for each metric.

TABLE I

SUCCESSFUL PREDICTION RATES BY INFORMATION SOURCE

Metrics	Answer type	
	USER-ACCEPTED	COMMUNITY-BASED
ASR and RSR	34.1%	30.3%
PR	33.7%	29.9%
Combined	34.5%	30.6%

TABLE II

SPEARMAN CORRELATION VALUES BETWEEN METRICS

Metric 1	Metric 2	Answer Type	
		USER-ACCEPTED	COMMUNITY-BASED
ASR	RSR	-0.03	-0.01
ASR	PR	0.52	0.55
RSR	PR	0.37	0.37

A. Rate of successful prediction

Table I shows the successful prediction rate using the different metrics separated (by information source) and combined. Interestingly we notice how combining the metrics does not improve the results substantially, for this reason we computed the correlation between the metrics.

Table II displays the Spearman correlation values between metrics for posts. We determined the correlation for posts rather than for tags as this provides us with a better insight in similarities between posts. While both ASR and RSR show moderate correlation with PR, they don't show any correlation with each other.

Overall, the metrics seem to be better suited to predict USER-ACCEPTED answers than COMMUNITY-BASED ones, possibly due to larger answering time distributions in COMMUNITY-BASED answers.

B. Distance between actual and predicted time bins

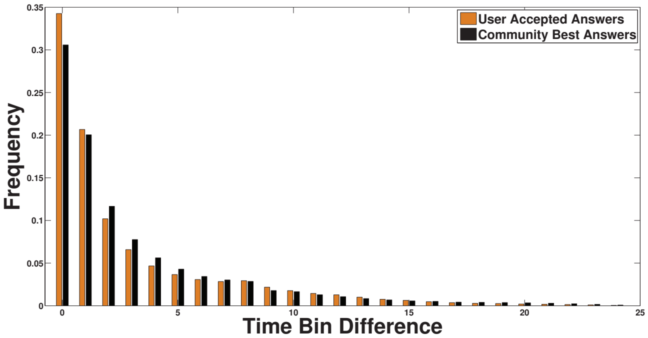

Fig. 1. Time Bin Differences for combined metrics

Figure 1 displays the absolute distance between the actual time bin and the predicted time bins for the User-Accepted and Community-Best Answers, both with combined metrics. The first group of bars corresponds to the successful prediction rate mentioned in Section IV-A. The graph also shows that 50% is classified within 1 bin from the actual bin, and 60% is placed within 2 bins. The distribution suggests that the algorithm performs reasonably well. However, due to the varying time ranges of the bins, the bin difference itself does not give an indication of the actual error.

C. Relative error

To obtain insight in the actual error we calculate the median relative error. As with the success rate we calculate these for the 3 distinct metric types for both the answer types. For the 'User-Accepted Answers' the median relative error is 3.12 for the user-based metrics, 3.33 for the tag-based metrics and 3.00 for the combined metrics. As for the 'Community-Best Answers' the respective errors are 3.45, 3.83 and 3.33. These results show that, as with the Successful Classification Rate, results on 'USER-ACCEPTED Answer' are better than on the 'COMMUNITY-BASED Answers'.

Based on these findings, we investigate further the relative errors of the 'USER-ACCEPTED Answer' and 'COMMUNITY-BASED Answer' datasets with the combined metrics.

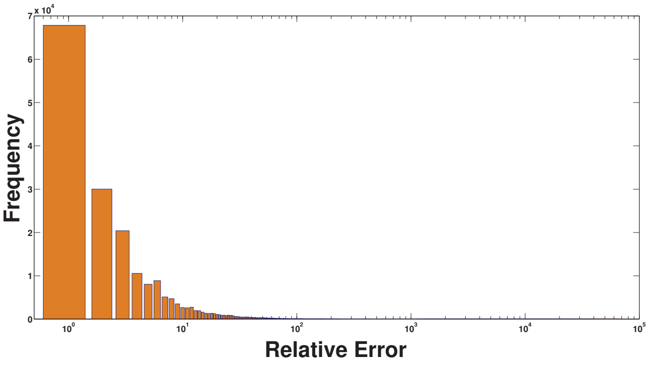

Fig. 2. Relative Error Distribution for USER-ACCEPTED Answers

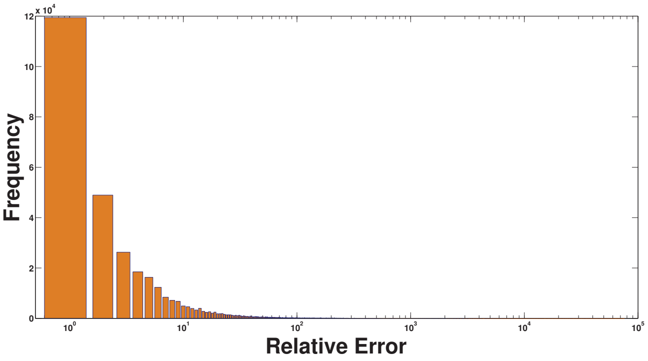

Fig. 3. Relative Error Distribution for COMMUNITY-BASED Answers

Figure 2 and Figure 3 show the relative error distribution for the User-Accepted Answers and Community-Best Answers respectively. Both plots show similar distributions, a large peak at 1, corresponding to relative errors below 1, followed by a steadily decreasing frequency as the relative error increases. The figures also show that over 60% of the predictions for both the User-Accepted and Community-Best Answers are acceptable, considering a relative error of 6 (600%) is acceptable, as assumed by Arunapuram et al [1].

V. Discussion

A. Recommendations and implications

During our investigation and by analyzing our results, we identified aspects that deserve further investigation:

Predicting Community-Based answers: Since the metrics we used for the prediction are based on tags and their popularity (measured from different angles: user and overall SO), we expected them to be better predictors for COMMUNITY-BASED answers rather than USER-ACCEPTED ones. Data proved us wrong. We did not investigate this aspect further, but studies can be designed and carried out to better understand the reason why USER-ACCEPTED answers seem more influenced by tags.

Tags: Due to our decision to remove unrepresented tags, if our approach was to be used in practice, it could result in worse predictions, which could lead to users being less satisfied. This aspect requires further investigation. Another aspect that we did not investigate further, but could have impact on the prediction, is the individual influence of each tag. We consider this an interesting avenue for future work.

Beyond tags: Bhat *et al.* report other post features that influence answering times [3]. Before any solid predictions can be made, these, together with the elements that lead to unanswered questions, will have to be further investigated. Moreover, an indication purely based on tags could negatively induce to only focus on tags rather than post quality to improve the ETA.

Major Programming Languages: Upon manual inspection we found that posts with a major programming language added as a tag (*e.g.*, Java) generally have a lower answering time. This effect could be due to the larger pool of subscribers to these programming languages, but other factors could be involved as well. Bhat *et al.* discovered a similar pattern researching the presence of a top 1,000 tag [3]. Understanding how different types of language tags work could help making better answering time predictions.

Acceptable Error: While we assumed an error of 600% to be acceptable, as also assumed by Arunapuram *et al.* [1], this might be an important aspect to research further. Knowing how much time askers are willing to wait longer than the predicted time can provide a clearer indication of when a prediction is good enough. This, in turn, can be used to determine the kind of resources a company like Stack Overflow would have to use to be able to provide predictions on answering times.

B. Threats to validity

Our results are endangered by threats to their validity, concerning both internal and external validity.

Internal threats: Taking a subset of all the questions on SO can internally threaten the validity of our research: With a biased subset (*e.g.*, having little variation in answering times) the k-NN algorithm could provide unreliable results. We tried mitigating this by taking a subset that seemed representative of the complete data set, both in tags and posts, having researched 5% of all the questions, representing 90% of all SO tags (before filtering underrepresented tags). It is possible that even though this subset seems to represent the whole data set it is still biased, as such a selection is subjective to human judgment (*i.e.*, what is considered representative).

External threats: An external threat to validity is the user base. Any prediction done depends on the assumption that the user base will continue to function in the same manner as it did during the investigated time span. To mitigate individual influence and predict human behaviour, we only selected well-represented tags. This does not entirely remove the human factor, but removes the influence that a small group (<15 people) can exert on the result.

Other sources of external threats are the algorithms that we used. It could be that the k-NN algorithm and the k-means algorithm are biased in a particular direction, creating systematic bias. While we cannot remove this systematic bias, we can reduce the non-systematic validity threats originating from the algorithms. One way we attempted to do this was by running each algorithm multiple times and comparing the results. The k-NN algorithm proved to be consistent over the different runs, suggesting that the bias from the data was not very influential. Also using a sampling rate of 0.1 for the training set reduced chances of bias originating from the training set selection. For the k-means algorithm, the comparison between runs had to be done by hand, making it subjective. We countered this subjectivity by having 3 independent people run the algorithm, select a set that they considered representative of the data, and then comparing the 3 k-means results. Due to the similarity between these results 1 was randomly chosen.

VI. Conclusion

As Q&A sites become increasingly popular, having an ETA as the question is created would be beneficial. We investigated a large dataset to see whether we can accurately predict answering times based on tags, computing metrics from different information sources. Of all post analyzed, 30-35% of the posts were predicted correctly, while the rest was predicted with a median relative error of 3.00, thus providing a reasonable ETA.

References

[1] P. Arunapuram, J. W. Bartel, and P. Dewan. Distribution, correlation and prediction of response times in stack overflow. In *Proceedings of CollaborateCom 2014*, pages 378–387. IEEE, 2014.

[2] M. Asaduzzaman, A. S. Mashiyat, C. K. Roy, and K. A. Schneider. Answering questions about unanswered questions of stack overflow. In *Proceedings of MSR 2013*, pages 97–100. IEEE Press, 2013.

[3] V. Bhat, A. Gokhale, R. Jadhav, J. Pudipeddi, and L. Akoglu. Min (e) d your tags: Analysis of question response time in stackoverflow. In *Proceedings of ASONAM 2014*, pages 328–335. IEEE, 2014.

[4] A. Bosu, C. S. Corley, D. Heaton, D. Chatterji, J. C. Carver, and N. A. Kraft. Building reputation in stackoverflow: an empirical investigation. In *Proceedings of MSR 2013*, pages 89–92. IEEE Press, 2013.

[5] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann. Design lessons from the fastest Q&A site in the west. In *Proceedings of CHI 2011*, pages 2857–2866. ACM, 2011.

[6] A. Rechavi and S. Rafaeli. Not all is gold that glitters: Response time & satisfaction rates in yahoo! answers. In *Proceedings of PASSAT 2011 and SocialCom 2011*, pages 904–909. IEEE, 2011.

[7] A. T. T. Ying. Mining challenge 2015: Comparing and combining different information sources on the stack overflow data set. In *Proceedings of MSR 2015*, page to appear, 2015.