Intuition vs. Truth: Evaluation of Common Myths about StackOverflow Posts

Verena Honsel, Steffen Herbold, Jens Grabowski
Institute of Computer Science, Georg-August-University Göttingen, Germany
{vhonsel,herbold,grabowski}@cs.uni-goettingen.de

***Abstract*—Posting and answering questions on StackOverflow (SO) is everyday business for many developers. We asked a group of developers what they expect to be true about questions and answers on SO. Most of their expectations were related to the likelihood of getting an answer or to voting behavior. From their comments, we formulated nine myths that they think are true about the platform. Then, we proceeded to use rather simple methods from statistics to check if these myths are supported by the data in the SO dump provided. Through our analysis, we determined that there is an effect for eight of the nine myths the developers believed in. However, for only four of the myths the effect size is large enough to actually make a difference. Hence, we could bust five myths the developers believed in.**

I. Introduction

Since StackOverflow (SO)[1] is a platform for asking and answering questions regarding programming, most developers use it during their everyday work. Through their work with the platform, developers get an intuition about postings, e.g., which questions are likely to be answered or what good and accepted answers look like. Our aim for this study within the MSR mining challenge 2015[2] is to evaluate if the intuition of the developers is correct or if they believe in myths. We sat down with a group of five developers working at our institute and performed a brainstorming. From the discussion within this brainstorming, we distilled nine myths, in which all five of the developers believed:

M1: Users with a high reputation are more likely to get an answer.
M2: Questions with source code are more likely to get an answer.
M3: Questions posted in American business hours are more likely to get an answer.
M4: Correctly capitalized questions are more likely to get an answer.
M5: Positively voted questions are more likely to get an answer.
M6: Questions that have duplicates are more likely to get an answer.
M7: Answers with source code get better votes.
M8: Answers with too much text get worse votes.
M9: New users violate rules more often.

The data provided in the mining challenge is a complete dump of SO, including the text of the posts, voting behavior, accepted answers, and information about the users. For the evaluation of our myths, we exploited all these different information sources. Due to the diverse nature of the myths and the different kinds of data, we decided to use a very simple approach for analysis. We define a property that describes the myth (e.g., high reputation) and then evaluate the impact it should have according to the myth. For example, developers believe that high reputation increases the likelihood that a question gets an answer. Hence, we split the data based on the reputation and evaluate the mean value of the answered questions for both splits. This very simple and versatile strategy was applied for the evaluation of all nine myths.

The remainder of this paper is structured as follows. In Section II, we outline how we handled the complexity of the provided data and facilitated our later analysis. Then, we describe our analysis approach in detail in Section III. Afterwards, we evaluate the myths in Section IV and discuss the threats to the validity of our results in Section V. Finally, we conclude the paper in Section VI.

II. Data

The data for the challenge was the XML dump of the SO content made available by Stack Exchange on September 26th, 2014. The size of the XML files once unzipped was about 89.96 Gigabyte (GB). To be able to handle the data complexity in a convenient way and to impose a structure suitable for our analysis, we created a MySQL[3] database from the provided XML dump and then used SQL queries to impose a structure suitable for our analysis. Concretely, we created a new table `PostsAttributes` which contains metadata about each posting, e.g., the timestamp, reputation of the user, type of the posting, etc. We extended the `PostsAttributes` with a small Java program that checked the correct capitalization of posts (see Section IV, myth *M4*). We then exported the metadata as Comma Separated Value (CSV) files. This way, we reduced the data for the analysis to two CSV files: one with metadata of the 7,990,787 questions that was about 1.17 GB in size, and one with metadata of the 13,684,117 answers that was about 1.61 GB in size. These files were small enough that we could evaluate them with RStudio[4] on a laptop with 8 GB of memory. The preprocessed data and the R script we used for the analysis are available online.[5]

[1] http://stackoverflow.com/
[2] http://2015.msrconf.org/challenge.php
[3] http://www.mysql.com/
[4] http://www.rstudio.com/
[5] http://bit.ly/1LRonqz

978-0-7695-5594-2/15 $31.00 © 2015 IEEE
DOI 10.1109/MSR.2015.58

III. Analysis Method

To evaluate the myths, we formulate them in a structured way and either accept or reject them based on their support in the data. The structure is based on two properties A and B, and according to the myth there should be a correlation between the two properties. Once we defined A and B for a myth, we split the data into two partitions based on property A: $data_A$ where A is fulfilled and $data_{\bar{A}}$ where A is not fulfilled. We then report the mean value of property B on both sets $data_A$ and $data_{\bar{A}}$. If the property is defined using some threshold t, e.g., $A = \textit{valueOf}(X) > t$, we calculate the mean of X. Moreover, we perform a Mann-Whitney-U test [1] to evaluate if the difference we observe is significant.

While this concept may seem abstract, when applied it is actually quite simple. Consider the myth "*M1*: Users with a high reputation are more likely to get an answer". Then our property $A = \textit{reputation} > t$ with an appropriate threshold t, and our property $B = \textit{hasAnswer}$. Our aim is to observe if there actually is such a correlation, i.e., if a high reputation means a higher likelihood of getting an answer. We start by splitting the data in two sets: one with the questions posted by users with high reputation and one with the questions posted by users with low reputation. Then, we evaluate the mean value of the attribute *hasAnswer* in both of these sets. Note, that *hasAnswer* is a binary attribute, i.e., the mean value is actually the percentage of answered question. Furthermore, we calculate the *p-value* for the significance of the difference between the mean values with the Mann-Whitney-U test.

We report all results within our case study and draw a conclusion from them. Our conclusions are based on two criteria:

- The results must be statistically significant with a significance level of 0.001, i.e., if we have 99.9% confidence in the difference. Hence, we accept the results as valid if the *p-value* < 0.001.
- For the myths *M1–M6*, we evaluate the percentage of answers given. In order to take the effect size into account, we only accept those myths as true where there is a difference of at least 10%. The other myths are busted.
- The myths *M7–M9* consider voting behavior and rule violations. We do not define an absolute effect size for these myths, because we expect the average number of votes/violations to be rather low. Instead, we define a proportional effect size and accept the myth as true if there is a change of at least 50% in the measured value if property A is fulfilled, in comparison to where property A is not fulfilled. For example, if the value where property A is not fulfilled is 1 and we expect an increase due to property A, we accept the myth if and only if the value with property A fulfilled is 1.5.

IV. Nine Myths about SO

In the following, we explain the rational for each myth, define the criteria used for evaluation of the myth , and report the results of the evaluation. We do not report the *p-value* for the evaluations. Due to the huge amount of data, all the effects we found are statistically significant with a *p-value* $< 10^{-16}$. A summary of the results is depicted in Table I.

M1: Users with a high reputation are more likely to get an answer

Rational: Reputation is an important measure of how trustworthy a user is. Users earn reputation, e.g., by getting upvotes and accepted answers. They can also loose reputation, e.g., by getting a downvote. Therefore, reputation is seen as an indicator that a user asks important questions and gives high quality answers. From this follows the intuition that questions asked by users with a high reputation are more likely to get an answer.

Criteria applied: This myth is already used in the example in Section III. We determined the threshold t for *high* user reputation as the upper quartile of the reputation of all SO users, which is 20, i.e., our property A is *reputation* > 20.

Result: There is a strong impact of the reputation on the the likelihood of answers. Users with high reputation get an accepted answer in 64.8% of the cases, whereas those with a low reputation only get answers in 30.7% of the cases. The effect size is with 34.1% quite large. Therefore, we **accept this myth as true**. Reputation is an important factor for getting answers. A similar result is stated by Movshovitz-Attias et al. [2], who retrieve patterns based on active question asking/answering and high user reputation.

M2: Questions with source code are more likely to get an answer

Rational: Our underlying assumption is that code examples within the post are conductive for the understanding of potential readers. A deeper understanding translates to more confidence in formulating an answer.

Criteria applied: The property A we evaluate here is if a question contains a `<code>` tag, i.e., part of the question is highlighted as source code. The property B is *hasAnswer*, as in the example used in Section III.

Result: There is only a weak impact of having source code in the question on the likelihood of getting an answer. Questions with source code get an answer in 59.8% of all cases, questions without code only in 52.3% of all cases. The effect size is rather small with only 7.5% and, therefore, too small to make a big difference. Therefore, this **myth is busted**.

M3: Questions posted in American business hours are more likely to get an answer

Rational: We expect that questions that do not get attention rather shortly after being asked to be pushed back in favor of newly incoming questions. Hence, we expect that questions posed in times when active users of SO are available are more likely to attract their intention and from this are more likely to get answered. Due to the English language of the platform and the huge number of developers working the the USA, we suspect a big impact of the American business hours.

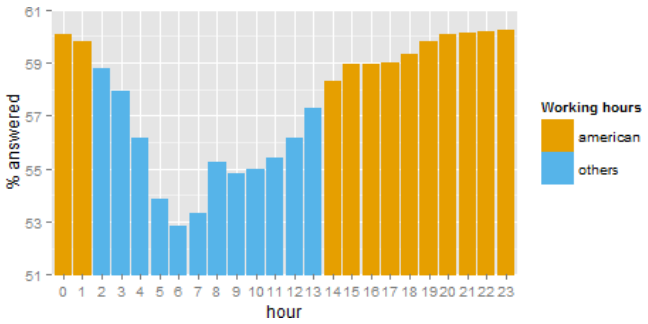

Fig. 1: Percentages of answered questions per hour.

Criteria applied: Property A are the American business hours, which we defined as 9:00 o'clock eastern time (UTC-05:00) until 17:00 o'clock pacific time (UTC-08:00), i.e. from 14:00 o'clock UTC till 2:00 o'clock UTC. We did not account for daylights savings time, which could have caused some noise at the boundaries. The property B is *hasAnswer*, as in the example used in Section III.

Result: There is only a weak impact of the American business hours on the likelihood of getting an answer. Questions asked within the American business hours get answered 59.4% of the time, whereas questions asked at other times only get an answer 55.4% of the time. Figure 1 visualizes this drop outside of the business hours. This finding is similar to Bosu et al. [3], where the authors also observe the same drop in efficiency of answering. However, we only observe an effect size of 4%, which is too small for us to accept this myth. Therefore, this **myth is busted**.

M4: Correctly capitalized questions are more likely to get an answer

Rational: There are two reasons for assuming that capitalization plays an important role. First, the readability and second, we assume that developers tend to answer more likely to posts where the authors exert themselves for writing and describing their problems properly.

Criteria applied: We applied a simple heuristic to check if capitalization is used correctly. First, we parsed the question with an HTML parser. Then, we removed all links and source code. For the remainder of the text, we heuristically determined the start of a sentence as (a) the start of the post or (b) a dot followed by a space. We then checked if the next character was upper case and use the percentage of correctly capitalized words as $capitalizationScore$. Because our heuristic is not perfect, we assume correct capitalization if $capitalizationScore > 0.5$, which is our property A. The property B is *hasAnswer*, as in the example used in Section III.

Result: There is a weak impact of correct capitalization on the likelihood of getting an answer. Correctly capitalized questions get answered 57.9% of the time, whereas questions without correct capitalization get an answer 52.7% of the time. The effect size is very small with only 5.2%. Therefore, this **myth is busted**. However, we observed that 92.4% of the asked questions use correct capitalization. Hence, it seems that this question is for the most part actually irrelevant, since correct capitalization is used anyway.

M5: Positively voted questions are more likely to get an answer

Rational: This myth is tangible, because if many people have same issues, they tend to vote the question targeted to resolve these. Because of this attention they are also more likely to get an answer.

Criteria applied: Our property A for this question is $votes = positiveVotes - negativeVotes > 0$, i.e., we say that a question is positively voted if it has more positive votes than negative votes. We count upvotes and marking as favorite as positive votes and downvotes as negative votes. The property B is *hasAnswer*, as in the example used in Section III.

Result: There is a strong impact of positive votes on the likelihood of getting an answer. Questions with more positive than negative votes get answered 66.2% of the time, whereas other questions only get answered 48.5% of the time. The effect size is quite large with 17.7%. Therefore, we **accept this myth as true**.

M6: Questions that have duplicates are more likely to get an answer

Rational: Similar to *M5* the post under investigation are of interest to a larger amount of people, so we assume them to get answered more probably.

Criteria applied: Our property A is if a question has a marked duplicate in the dump. The property B is *hasAnswer*, as in the example used in Section III.

Result: There is a strong impact of duplicates on the likelihood of getting an answer. Questions with marked duplicates get answered 77.6% of the time, whereas other questions only get answered 57.3% of the time. The effect size is quite large with 20.3%. Therefore, we **accept this myth as true**. However, we observe that only 1.3% of the questions have duplicates. So while this effect is quite strong, luckily, there are still not too many duplicates.

M7: Answers with source code get better votes

Rational: From the point of a developer struggling with a certain programming issue it can be of great help to get example code or workarounds, which helps to solve it. Thus, we think that answers containing source code get more positive votes.

Criteria applied: Our property A is the same as for *M2* but this time defined for answers, i.e., if an answer contains a `<code>` tag. We use $votes = positiveVotes - negativeVotes$ (see *M5*) as foundation for property B.

Result: There is an impact of having source code within the answer on the voting behavior. Answers with source code get 2.45 votes on average, whereas answers without code only get 1.69 votes. This is an increase of 45%, which is below the 50% increase threshold for the effect size. Therefore, this **myth is busted**. Source code gives an advantage, but not a major one.

M8: Answers with too much text get worse votes

Rational: This myth also deals with the question, which attributes a *good* answer should have. Contrary to *M7*, where we state involved source code as positive, we rank too much text as negative, because we assume well explained, but not overdrawn answers to be the most satisfactory for the reader.

Criteria applied: We use the *length* of an answer in characters as foundation for our property A. We determined the threshold t for *too much text* as the upper quartile of the length, i.e., 897, our property A is $length > 897$. We use $votes = positiveVotes - negativeVotes$ as property B, same as for *M7*.

Result: There is a strong impact of the length of the answer on the voting behavior. Long answers get 3.03 votes on average, whereas other answers only get 1.92 votes. This is an increase of 58%, which is above the 50% increase threshold for the effect size. However, the effect is the complete opposite of what we expected, since we were actually expecting a decrease. Users tend to favor rather long answers. Therefore, this **myth is busted**. Instead, the opposite is true and long and detailed explanations actually translate to more positive votes.

M9: New users violate rules more often

Rational: For our last myth we assume unexperienced users to violate behavioral rules more likely. One explanation is that they are not aware of the restrictions, which mean, e.g., offensive text. Another could be that they register on purpose to spread spam.

Criteria applied: For the evaluation of this question, we use both questions and answers. As foundation for property A, we use the *accountAge*, which we define as the difference between the user creation date and the date of the posting to determine. We define a user as new, if the *accountAge* is less than one week. For property B we use the number of $violationVotes > 0$, i.e., we are interested in any rule violation. We define *violationVotes* as the number of offensive, closed, and spam votes.

Result: There is a strong impact of the user account age on rule violations. New users violate rules in 1.0% of the cases, whereas users that have been on SO for at least one week only violate rules in 0.4% of the times. This is an increase of 150%. Therefore, we **accept this myth as true**. However, only one out of 100 questions from new users violates rules, which are still very few violations considering the number of users and postings on SO.

V. Threats to Validity

There are two major threats to the validity of our results.

- Most of our myths are related to problems associated with programming. However, not all questions asked on SO are related to programming, there are also other topics. Our findings regarding those myths might change if we remove posts unrelated to programming topics. However, since programming related posts are the overwhelming majority, we do not expect major changes due to this.

M1 (true)	
answered questions if the user reputation is high	64.8%
answered questions the user reputation is low	30.7%
M2 (busted)	
answered questions with code	59.8%
answered questions without code	52.3%
M3 (busted)	
answered questions asked in American business hours	59.4%
answered questions asked outside of American business hours	55.4%
M4 (busted)	
answered questions with correct capitalization	57.9%
answered questions with wrong capitalization	52.7%
M5 (true)	
answered questions if question highly voted	66.2%
answered questions if question not highly voted	48.5%
M6 (true)	
answered questions if there is a duplicate	77.6%
answered questions if there are no duplicates	57.3%
M7 (busted)	
mean number of votes for answers with source code	2.45
mean number of votes for answers without source code	1.69
M8 (busted)	
mean number of votes for long answers	3.03
mean number of votes for answers without much text	1.92
M9 (true)	
mean number of violations from new users	1.0%
mean number of violations from old users	0.4%

TABLE I: Summary of the results.

- The threshold for the effect sizes were determined based on our experience and our intuition. The conclusions which myths are valid and which should be rejected depend on these thresholds, which means that if our choices were bad, the results would be invalid.

VI. Conclusion

Our study shows that developers have a remarkably good intution regarding postings on SO. For eight of the nine myths we found a statistically significant effect for what the developers expected. Only in case of the relationship between answer length and voting behavior was the rational at odds with what we observed in the data. However, our study also shows that the developers overestimated the effect of the myths. Only in four cases the effects are sufficiently large to make a noticable difference. Hence, we could bust five myths, because they don't actually make a big difference or, in one case, are actually completely wrong.

Acknowledgements

We would like to thank Gunnar Krull for the technical and administrative support that enabled us to handle the large amount of data.

References

[1] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Ann. of Math. Stat.*, vol. 18, no. 1, pp. pp. 50–60, 1947.

[2] D. Movshovitz-Attias, Y. Movshovitz-Attias, P. Steenkiste, and C. Faloutsos, "Analysis of the Reputation System and User Contributions on a Question Answering Website: StackOverflow," in *Proc. 2013 IEEE/ACM Int. Conf. on Advances in Social Netw. Anal. and Mining*. ACM, 2013.

[3] A. Bosu, C. S. Corley, D. Heaton, D. Chatterji, J. C. Carver, and N. A. Kraft, "Building Reputation in StackOverflow: An Empirical Investigation," in *Proc. 10th Working Conf. on Mining Softw. Repositories (MSR)*. IEEE Computer Society, 2013.