SE MID SEM EXAM

Yash Patel
Lakehead University
pately145@lakeheadu.ca

***Abstract*—This paper is aimed at analysis of question answer site 'Stack Overflow'. This analysis is aimed at finding out why stack overflow achieved and maintained its reputation in programming question and answer forums. Original Study done by [1] showed that median time for a question to receive its first response is 11 minutes. This analysis will provide a deeper understanding of discrepancy of the expert user and new users along with analysis of its working over the years from 2008 to 2010.**

I. Introduction

With increasing popularity of online forums for advises, question and suggestion, came a new forum 'Stack Overflow'. This forum is specifically designed for programmers to seek advise from other programmers. It receive huge success being technical forum for programmers. [] studied the open source data made available by Stack overflow to come to some conclusions and understanding of the factors affecting the success of stack overflow. Initial studies show that Stack Overflow attracted traffic of over 7 million on monthly bases. This can be attributed to many factors out of which one which stood out the most was that it had answer rate of over 90

This study aims at analysis the data provided by Stack Overflow and study it based on certain criteria.

II. Brief working of Stack Overflow

As its a forum, it has all the basic things of question and answer forums, but in addition to this it also has tag and snippet system for facilitating programming based questions. Here user can ask question along with code snippet to other users and other user can reply in the same manner and provide explanation and solution snippet with it. Other users can agree as well as disagree with the suggestion of other user. They can do so by commenting as well as downvoting or upvoting answers and questions. The thing that makes Stack Overflow different from other platforms is that it has reputation based system. This allows users to have some credibility to what they are saying or asking in the forum.

Purpose of this study is to analyse the data to arrive at certain conclusion and understanding of the working of Stack Overflow and why it has achieved what it has achieved.

III. Question 1 : What is the distribution of answers and thread lengths for all questions?

As explained in the working of Stack Overflow, each question has multiple answers and comments. The total of these questions and answers is known as thread length. So

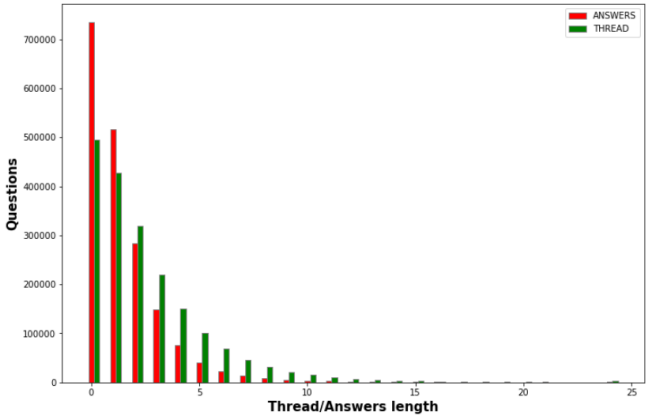

Fig. 1. Question 1: Answer and Thread length for all questions

analyze the thread length, I've done summation of number of comments and number of answers for that question. Doing so will give thread length for that question. I've plotted thread and answer length of questions with number of questions. Here in fig 1, Y axis signifies the number of questions have thread and answer length corresponding to the X axis value. Here we can see that most of the questions have answer length ranging from 0-3, meaning picking any question at random will have very high chance of it have 0-3 answers. Similarly most of the questions have thread length ranging from 0-5, meaning picking any question at random will have high chances of it having thread length ranging from 0 to 5. Further it can be inferred that most the answers provided by users are credible to the extent that further answering and commenting on the answer is not required in most of the cases. This can easily be the main reason behind success of Stack Overflow where users don't have to refer to multiple answers to arrive at a conclusion.

IV. Question 2: How long do users wait to receive answers?

One of the major factors that would affect the success of the Stack Overflow is the response time, meaning the wait time for answer. To study this, I've plotted graph which shows first answer time of the question as a percentage of all the questions. Red line in the graphs stands for this first answer time. It shows that more than 60 percent of questions receive first answer within the first 20 minutes of the question being posted, with 90 percent of the questions being answered within

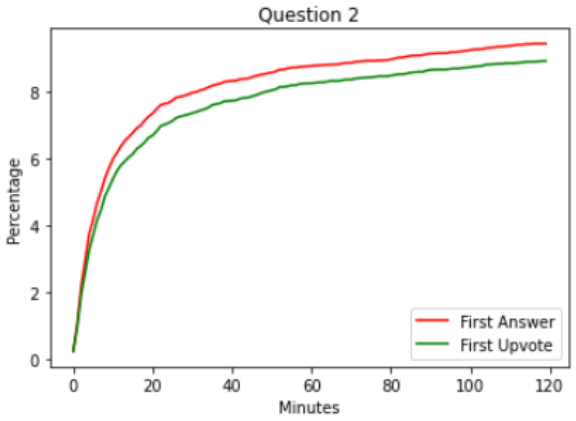

Fig. 2. Question 2: Answers in 2 hours after questions are posted

first 2 hours of the question being posted. Green line in the graphs does the same but it shows the first upvote received by question as a percent of all the question with time. From upvotes it can be seen that 80 percent of posts receive their first upvot wihin first hour of the question being posted. It can be inferred from this that within first hour a reliable solution might be present for the question as their is high chance of upvot for the answers. From this it can be said that response time of questions is around 1 hour, meaning there is high probability of solution being presented with first hour of its posting.

V. Question 3: How did this rapid answer time emerge over the site's history?

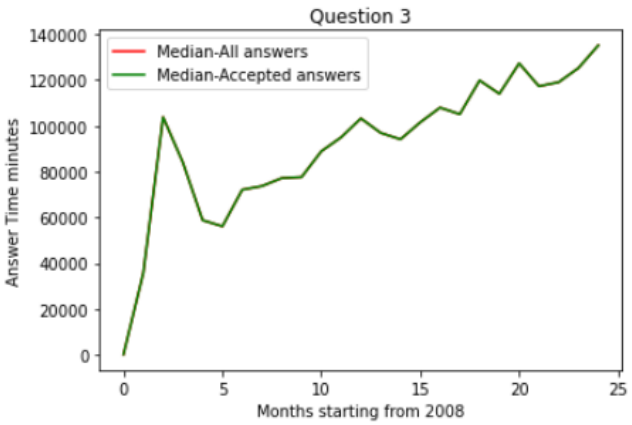

Fig. 3. Question 3: Answer time over the months

This study will explore the history of response time of Stack Overflow. To achieve this median of all answer time is calculated per month and plotted in historical format. Similarly median of accepted answer time is also calculated per month and plotted in historical format. Here in the graph both red and green lines are overlapping to a great extent, signifying that median of answer time and median of accepted answer time is same for most the months. Additionally it can be seen that medians are increasing with passing months. From this it can be said that users are taking more time to come to answers or the frequency of users visiting and answering question is decreasing a bit each month. From month number 15 it can be seen that median change is a lot less between months as compared to earlier months signifying that we might be able to get plateau in coming years and find the average median time to answer a question.

VI. Question 4: Which tags are associated with first and slow response time?

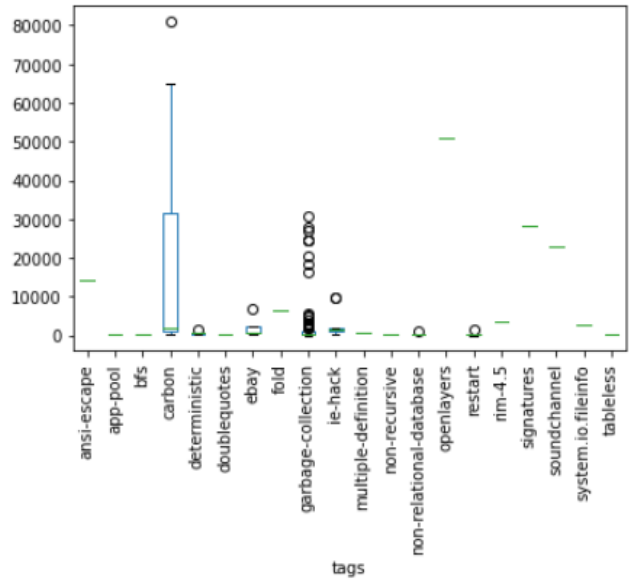

Fig. 4. Question 4: Median of fast question tags

This is a study for response time based on different tags. This is done using median times of tags and categorizing them into slow and fast tags based on the median value. Slow tags are tags which have median response time greater then 12 hours while fast tags are tags which have median time less the 10 minuts. After these top 20 in both categories are selected to be plotted. Based on the graphs observed it can be said that ansi-escape,app-pool and bfs have fastest response time in fast category while 3-way-merge, bmp, and code-behind have slow response time slow category. It can be inferred that tags associated with fast response time are buzz words and are much more popularly used amongst the programmers as compared to the slow tags. This might also imply that slow tags have less number of users who have knowledge about that tag or question might be to difficult for these tags.

VII. Question 5: How to calculate the reputation of users?

As discussed earlier in the working of Stack Overflow, reputation of user plays a major role. When a user with high reputation answers a question, most of the users will think

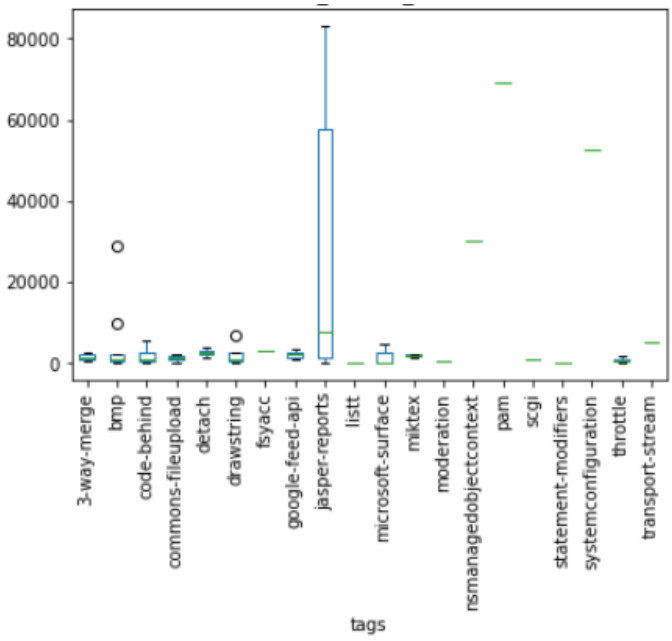

Fig. 5. Question 4: Median of slow question tags

```
In [32]: user=int(input("Enter the user ID:"))
         date=(input("Enter the date in proper format in proper format eg. 2010-02-19 ie.YYYY-MM-DD "))
         # print(reput(9,"2010-02-19"))
         print("Reputation of entered user is:",reput(user,date))

         Enter the user ID:9
         Enter the date in proper format in proper format eg. 2010-02-19 ie.YYYY-MM-DD 2010-02-19
         Reputation of entered user is: 1678
```

Fig. 6. Question 5: Input and output

that the answer provided will be more reliable as compared to answer provided by low reputed user. This allows narrowing down to answer in fast manner. Stack Overflow does this so by using complex algorithms to calculate users reputation based on his/her activities over the course of time, with there being upper limit for daily reputation gain to differentiate long term reputed users from the newly reputed users. For making of this function following criteria were used to calculate users reputation on a certain date.
Following were the criteria used to calculate reputation of a user:
-question is voted up: +10
-answer is voted up: +10
-answer is marked “accepted”: +15
-your question is voted down: 2
-your answer is voted down: 2
-daily limit of upvotes and downvotes aggregation to 200

Limitations of the function: This function has the following limitations:
-Not all rules were considered when calculating the reputation due to limitations on the data like bounty
-Need date to be inserted in proper format of YYYY-MM-DD, it won’t handle any other format and will through up an error if date format is not correct.

VIII. Summary

This analytic report studied Stock Overflow data and came to various conclusions with regard to the high success rate of Stack Overflow which includes high response time, reputation system, response time over months and response times per tags. This study gave a brief overview and understanding of some of the factors by which Stack Overflow is considered a good questiona and answer platform for programmers.

References

[1] Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G. and Hartmann, B., 2011, May. Design lessons from the fastest qa site in the west. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 2857-2866).