MidTerm Exam: *Design Lessons from fastest Q&A Site In West*

COMP-5413-WDE: Special Topics in Software Engineering

Name: Nilam Sivaji Bhosale
Student Id: 1138964
Department of Graduate Studies
MSc. Computer Science
Lakehead University.

ABSTRACT

Stack Overflow is a question and answer website for programmers, both professionals and hobbyists. This website intends to provide quick answers in the technical domain which helps in improving utility and performance in the field. This model helps the community to grow by building a public platform with a definitive collection of coding Q&A's allowing only questions which are very specific to a small focused problem. Stack Exchange is a network of 178 question and answer sites, including Stack Overflow, the largest and most trusted online community for developers to learn, share their knowledge, and advance their careers. This article will discuss the background, analysis based on various scenarios and related research for such Q&A websites and in particular about Stack overflow. This report summarizes the results of a data analysis conducted over three years(2008-2010) on Stack Overflow. I tried to analyze the Distribution of answers and thread length, response time for all tags associated with the questions and points which needed to be considered while calculating the reputation of users.

Keywords: Question and Answer, Reputation, Response Time, Stack Overflow, Posts, Datasets, comments

- INTRODUCTION

With the increase in usage and availability of the Internet these days it has become very easy to get some knowledge over the internet regarding any specific topic, with this it has also been observed that individuals are increasingly relying on peers for any information, advice, solution or help. As stated in [1] this practice of relying on peers has shown a positive result and hence more and more people are getting inclined towards it. So as a result, millions of people learn from the shared knowledge acquired by individuals from the places like community build encyclopedias – Wikipedia, discussion forums – Usenet, social networks, Q&A's sites-yahoo answers, stack overflow, etc. And such Q&A sites have grabbed the attention of researchers due to the increasing popularity and demand of such content to research the reasons for the success of this kind of website. Large and specialized organizations regularly integrate Q&A software with these websites, which is typically structured as a community that allows users in related disciplines to discuss and answer common and specialist queries [2].

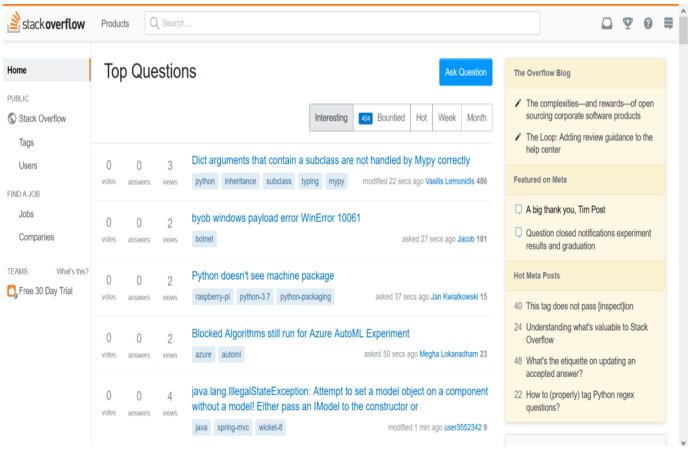

Figure 1 : Basic Overview of StackOverflow

This study examines elements in the design and evolution of Stack Overflow (SO), which is a very popular Q&A site for programmers and software engineers, and how they led to the site's success. Stack Overflow was first launched to public domain on 15th September 2008 with its beta version, and since then it has gained a huge popularity and has been evolving till date. Recently, in 2021 Stack Overflow was valued for $1.8 billion and was sold to South African media Company Naspers [3]. Stack Overflow has become one of the most visible places for expert knowledge sharing around software development in less than two years of its launch to public and it has a high response rate of over 90% and a median answer time of only 11 minutes, with over 300,000 registered users and >7 million monthly views (as of August 2010).

ENTITIES USED IN SO :

Posts :The term "post" refers to both queries and responses. Only questions, answers, and articles are subsets of posts.

Reputations: The community's faith in you is measured by your reputation. Posting good questions and useful replies is the most effective strategy to acquire reputation. Hence, Reputation of users is determined by the score of the post

Badges : Badges are small pieces of digital flair that you may acquire for nearly any type of Stack Overflow activity. Basically We can see the number and type of badges we

earned next to our username and reputation score around the site[4].

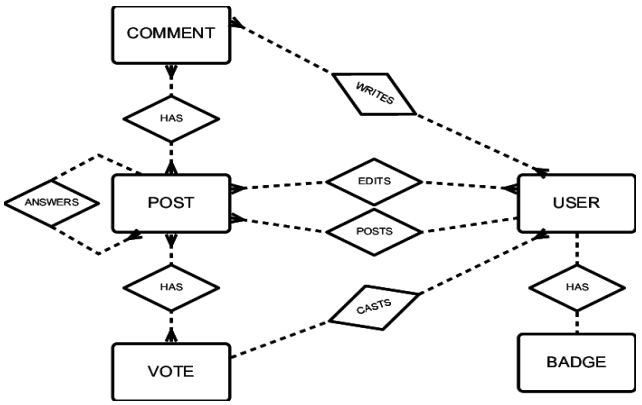

Figure 2 : Entities in SO

Tag : A tag is a word or phrase that groups together inquiries that are similar to yours. It's easier for people to locate and answer your question if you use the correct tags.

Comments - Comments"Post-It" notes left on a query or answer are known as comments. They can be highlighted and upvoted but not downvoted, but they do not earn reputation. There is no revision history, and if they are destroyed, they are permanently lost. We can always comment on our posts and any part of questions. Commenting on other people's postings, on the other hand, is a privilege earned through reputation[4].

Vote up - The community indicates which queries and answers are the most beneficial and suitable by voting them up. It is possible to vote up both postings (questions and answers) and comments.

RELATED WORK

Prior researchers have looked into such question answering websites (such as Stack Overflow and Yahoo! answer) and complementary techniques to knowledge sharing to understand the impact created in society and the reasons for their success. A lot of study has been done on Stack Overflow questions and answers. Its content has so far attracted a large number of researchers. Few of such researches have been discussed as under.

Ye, Xing, Z., & Kapre, N. (2016) present a paper where they focused on Why do developers share URLs on Stack Overflow? What kind of response has the community had to the knowledge that has been shared? What are the topological and semantic features of the resulting Stack Overflow knowledge network? Has this knowledge network reached a point of stability? If that's the case, how will it achieve stability? Answering these questions can assist the software engineering community better understand the knowledge diffusion process in programming-specific Q&A sites like Stack Overflow, allowing for more effective knowledge sharing, knowledge use, knowledge representation, and knowledge search. [6]

Barua, Thomas, S. W., & Hassan, A. E. (2012), In this study, they suggested analysing the text content of the posts—a rich and previously untapped source of information—to establish the overall topics of developer conversation, topic trends, and topic interaction patterns. Understanding these topics could help programming language and tool developers identify usage trends, commercial vendors analyse product adoption rates, and Q&A sites understand how their information content is being used[7]

Shah, Radford, Connaway, Choi, and Kitzie (2012) Using a grounded theory approach, they created a typology with four key categories: unclear, difficult, unsuitable, and many questions to identify causes for failures for these questions. They discovered that a large percentage of the rejected questions were either too complex or too wide, demonstrating the difficulty in constructing logical inquiries[8].

DATASET USED FOR ANALYSIS :

Question response times on Q&A sites have an impact on the longevity of online communities, as faster responses lead to higher user happiness and engagement. I looked at a large dataset from StackOverflow to see what factors were linked to response time and scenarios which are mentioned in Task. Under the Creative Commons Licence, Stack overflow makes its data freely available in XML format such as badges.xml, Comments.xml, Posts.xml, Users.xml, Posts.xml, votes.xml. In this report, I have majorly used posts.xml which contains post_type_id, answers_count, Creation_date, Parent_id and id of the user who created the post. Other tables which I have used for analysis are Post_hostory, Votes, Users and Comments. Also to understand some Data dependencies and relationship between tables, I have used Database schema documentation provided by Stackoverflow and StackExchange Data Explorer.

CHARACTERIZATION AND ANALYSIS OF STACKOVERFLOW DATASET

1. *DISTRIBUTION OF ANSWERS AND THREAD LENGTH OF ALL QUESTIONS*

In First of part of work, I focused on only the posts table as it contains all required data required to solve the question.

To extract the posts from the stack overflow dataset, I used posts.csv file from the dataset which contains all the posts (i..e., questions and answers and so on) on Stack Overflow.

I retrieved data based on post_type_id which has value 1 because we need to find out all answers and comments associated with that question. To get such values i filtered the data based on Post_type_id which has various attributes such as

1 = Question
2 = Answer
3 = Orphaned tag wiki
4 = Tag wiki excerpt
5 = Tag wiki and so on.

Each post can also have a list of comments associated with it. Comments are used for follow-up and statements of agreement or disagreement. I was able to calculate the thread count by summing up the total number of answers and Comments. Answers and comment count columns are already present in the Post table. As we know that one question can have multiple answers and comments so I performed groupBy operation on thread and answer count to get the results.

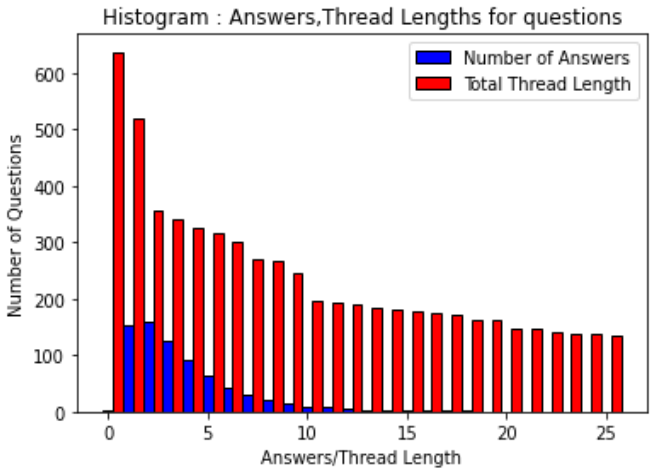

Figure 3: Answers, Thread Length for all questions

The complete distribution of answers and thread length is shown in Figure 2. and it's clear that, Around 200 questions have an average of 25 thread length (Answers and comments). I have calculated thread length only for 25 answers to fit the data into x and y axis.

HOW LONG DO USERS WAIT TO RECEIVE ANSWERS?

I have considered three different scenarios while solving this problem. 1) First answer is posted within 2 hours. 2) Answer which receives positive response. 3) Answer which is accepted by the questioner.

I will explain all scenarios in detail below:

A] **First answer is posted within 2 hours :**

From the above figure, it's clear that very few answers are posted within two hours. Average of 25% of questions are answered within 40 minutes. While calculating the above scenario, I have considered the posts table as all information related to questions and answers are present in the post dataset.

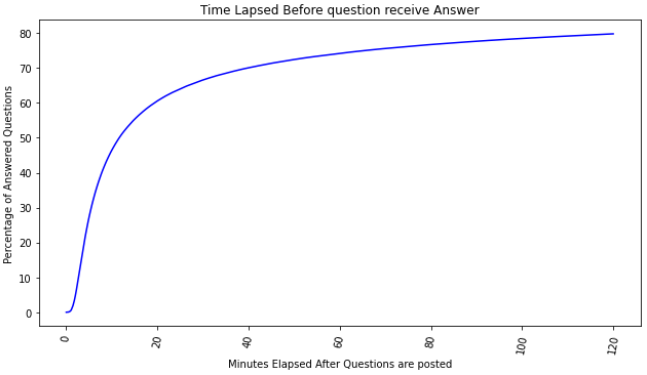

Figure 4 : Answers within 2 Hours after question posted

After that i have selected data where post_type_id = 1 and 2 where (1 = question, 2 = answers). Then I grouped together based on parent_id. Parent_id is present for that particular post only if PostTypeId=2. It will group all questions together. Calculated time for all answers to check whether those are answered within 2 hours timestamp for not. From Figure(4) It's clear that Approx 70% of questions were answered in around 70 minutes. The median time for upvoted answers is 25 minutes. We get the faster results for upvoted answers because upvoted answers are a subset of questions with questions with answers.

B] Answer which receives positive response.

To calculate the upvoted answers I have considered votes and posts tables in which VoteTypeId=2 indicates upvoted and PostTypeId=2 indicates the answer. Then performed a merge operation on both tables based on PostTypeId and Grouped together to get the combined answers. In the Vote table, creationDate contains only the date without timestamp so while retrieving the data i checked the answer which has received positive vote within 25 days. From the above figure, it's clear that the highest percentage of answered questions which were upvoted within an average period of five days.

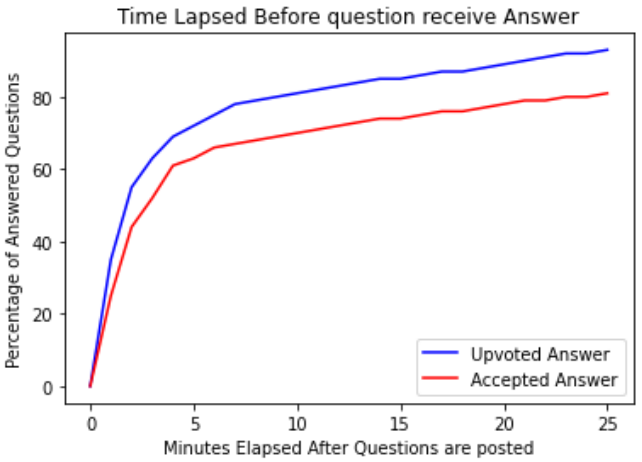

Figure 5 : Upvoted and Accepted Answers

C) Answers which are accepted by the questioner.

To calculate the Accepted answers I have considered votes and posts tables in which VoteTypeId=1 indicates that the question is accepted by the originator and PostTypeId=1 indicates the question. Then performed a merge operation on both tables based on AccpetedAnswerId(which is present only if PostTypeId=1) and postId.

Then, performed a group by operation to get the combined records based on the postId . Finally, Retrieved the answers which were accepted within 25 days and while calculating the daysI considered the date on which that answer was accepted by the questioner and date on which that question was generated.From the above figure, it's clear that approx 10% of answered questions were accepted.

For all three scenarios, I have taken a cumulative value to calculate the Time elapsed after questions are answered.

From Figure(5) It's clear that Approx 80% of questions were answered in around 20 days.

How did this rapid answer time emerge over the site's history?

I have considered below four different scenarios while calculating the answer time. 1) All Answers asked each month. 2) Accepted Answer 3)First Upvoted Answer 4)First Answer

1) **First Upvoted Answer** :

Median Time for upvoted answers has been around 15 minutes since summer 2009. It increased for some period of time but again it declined in 2009. To calculate the upvoted answers, Merged the same two tables based on postTypeId.

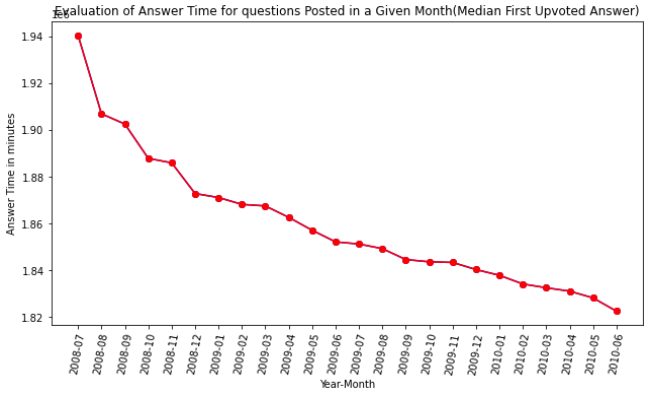

Figure 6: First Upvoted Answer

From the figure 6, it is observed that median times for Upvoted answers have started decreasing gradually after 2009.

1) **All Answers asked each month** :

Used Posts table to calculate all the answers within each month. Removed timestamp from creation date to retrieve year and month. Performed Group by operation on year and month and also calculated the answer time for all questions in minutes by taking average of Total answers.

2) **Accepted Answer** : Used Posts and votes table to calculate all accepted answers within each month.Checked for voteTypeId=2. Performed the same operations on date as mentioned in point (1). Merged Votes and Posts tables based on AcceptedAnswerId and Post Id to get the results.

3)First Answer :

Performed Group by operation on year and month and also calculated the answer time for all questions in minutes by taking average of Total answers.

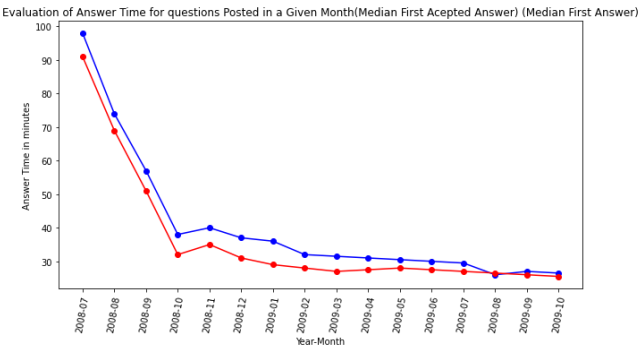

Figure 7 : Accepted and First Answer.

From the figure 7, it is observed that median times for all answers and for accepted answers have been around 25 minutes since 2009.

Which tags are associated with first and slow response time?

A tag in stackoverflow is a term or phrase that expresses the question's topic. By categorizing inquiries into distinct, well-defined categories, tags let professionals connect with queries they can answer. Tags can also be used to find questions that are relevant or interesting to us. When we click on a tag underneath a question, we will be sent to a page that lists all of the questions that fall under that tag. On Stack Overflow, we also get a description of what the tag is and how it should be used. Each question may only contain 5 tags at a maximum.

I have used below steps to calculate the tags which are associated with all answered questions with fast and slow time.

-Retrieved the data based on PostTypeId=1 and PostTypeId=2 to calculate the time difference.

- Retrieved all the tags associated with the Id's and spitted them to check the Response time for each Id.

- Also Get the counts for each id (i.e. number of occurrences its appeared in table)

- Calculate the fast and slow response times for each tag.

How to calculate the reputation of users?

Users on Stack Overflow can earn reputation points. A user's reputation indicates how active they are on the site. Posting good questions and useful replies is the most effective strategy to build reputation. Basically How many votes you are getting on your posts cause us to gain or lose the reputation. We can gain reputation when our question/answer is voted, answers are awarded a bounty by the user offering the bounty, answers become accepted, when we accept an answer written by someone else. We can lose the reputation when one of the questions/ answers is voted down, an upvote on any one of the questions or answers is removed. When we accept our own answers it does not increase the reputation.

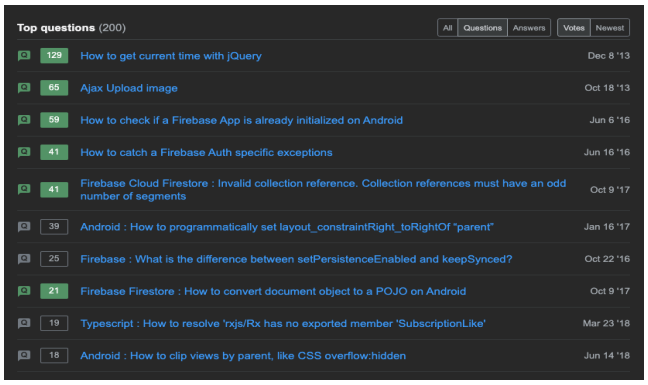

Figure 8 : Reputation of Users.

From Figure 8, As you can see from their top-voted questions and responses, this individual mostly built their reputation through asking questions.

Tables which I have considered while calculating the reputation of users are Posts, Votes, Users, PostHistory, Tags, comments.

Below are the steps which i have performed to calculate the reputation of user :

1. Loaded all the datasets (Posts, Users, Post_History, Votes,Comments)
2. Used UserId and Current Time as an input.
3. Performed Inner join on posts and votes table based on PostId and Id to get the common results for entered userId.
4. Considered below scenarios to calculate the reputation :

a) Total Reputation from accepts

- Checked the VoteTypeId=1. It means the answer is accepted by the originator. For that i have added 15 points in reputation.

b) ReputationFromAcceptedAnswers

- Retrieved data from posts table based on OwnerUserId And AcceptedAnswerId.
- OwnerUserId is nothing but UserId for which we are calculating the reputation.
- Apply null check on AcceptedUserId as it should not be null to calculate the reputation.

c) Reputation From Votes

- If PostTypeId is 1 (indicates Question) , VoteTypeId =2, CommunityOwnedDate is greater than creation date then 10 points given as Reputation
- If PostTypeId is 2 (indicates Answer) , VoteTypeId=2(Upvoted)Community OwnedDate is greater than creation date then 10 points given as Reputation
- IfVoteTypeId=3(Downvoted)and Community OwnedDate is greater than creation date then 2 points removed from reputation.
- So users will gain 10 points each, answer or question when any one of them is voted up.

4.Community OwnedDate present only if the post is community wiki'd.

5. After Considering above all four points above, I was able to calculate the 50 points for reputation. But the actual reputation for that particular user is given as 98 in Users table.

REFERENCES

[1] A. O. O'Donnell, "J. Learning from peers: Beyond the rhetoric of positive results," *Educ Psychol Rev 6, https://doi.org/10.1007/BF02213419,* p. 321–349 , 1994.

[2] Wikipdia, "Q&A Software," Wikipedia, 06 02 2022. [Online]. Available: https://en.wikipedia.org/wiki/Q%26A_software.

[3] B. Dummett, "Stack Overflow Sold to Tech Giant Prosus for $1.8 Billion," The wall street journal, 02 June 2021. [Online]. Available: https://www.wsj.com/articles/software-developer-community-stack-overflow-sold-to-tech-giant-prosus-for-1-8-billion-11622648400.

[4] "Stack Overflow," Stack Overflow, 2022. [Online]. Available: https://stackoverflow.com/.

[5] T. Ahmed and A. Srivastava, "Understanding and evaluating the behavior of technical users. A study of developer interaction at StackOverflow," SpringerLink, 27 February 2017. [Online]. Available: https://link.springer.com/article/10.1186/s13673-017-0091-8.

[6] D. Ye, Z. Xing and N. Kapre, "The structure and dynamics of knowledge network in domain-specific Q&A sites: a case study of stack overflow," SpringerLink, 19 April 2016. [Online]. Available: https://link.springer.com/article/10.1007/s10664-016-9430-z.

[7] A. Barua, S. W. Thoms and A. E. Hassan, "What are developers talking about? An analysis of topics and trends in Stack Overflow," 01 November 2012. [Online]. Available: https://link.springer.com/article/10.1007/s10664-012-9231-y.

[8] L. Zhe and B. J. Jansen, "Questioner or question: Predicting the response rate in social question and answering on Sina Weibo," Science Direct, March 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457317307343?casa_token=LPQERdIVcLoAAAAA:kVd24J3DCQCscylbGahlcJnIcKymM3p9ZvEh1wf4q4263KbskakMGEIB5vggk1UOqzhhgqqAdA.

[9] T. Santos, S. Walk, R. Kern, M. Strohmaier and D. Helic, "Activity Archetypes in Question-and-Answer (Q8A) Websites—A Study of 50 Stack Exchange Instances," ACM Digital Library, 1 March 2019. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3301612.

[10] Y. Yao, H. Tong, T. Xie, l. Akoglu, F. Xu and J. Lu, "Detecting high-quality posts in community question answering sites," Science Direct, 1 May 2015.

[Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025514011 89X?casa_token=ai8ay4Phi4AAAAAA:_EKuml74oR42JL2a72W1tGtxj3KafGjlVolv53IwXxXNO8Dpu6dH-SGZAPMxVeCZx8r2q_D4gg.

[11] D. Russo, P. Hanel, A. Seraphina and N. v. Berkel, "Predictors of well-being and productivity among software professionals during the COVID-19 pandemic – a longitudinal study," SpringerLink, 28 April 2021. [Online]. Available: https://link.springer.com/article/10.1007/s10664-021-09945-9.

[12] "How the R community creates and curates knowledge: an extended study of stack overflow and mailing lists," SpringerLink, 18 August 2017. [Online]. Available: https://link.springer.com/article/10.1007/s10664-017-9536-y.

[13] R. Verma and A. Srivastava, "A dynamic web service registry framework for mobile environments," SpringerLink, 05 January 2017. [Online]. Available: https://link.springer.com/article/10.1007/s12083-016-0540-6.

[14] J. Ahn, A. Pellicone and B. Butler, "Open badges for education: what are the implications at the intersection of open systems and badging?," Research in Learning Technology, 08 08 2014. [Online]. Available: https://journal.alt.ac.uk/index.php/rlt/article/view/1510.