STACK OVERFLOW BIG DATA ANALYSIS *

Tony Liu
College of Engineering
University of Miami

ABSTRACT

Stack Overflow (SO) is a community-based question answering service that targets developers and software engineers and has achieved significant success. This paper explored the characteristics of SO in terms of its users, questions and tags with data provided by Google BigQuery. The main findings of this work include the polarization of SO users, the question-answering pyramid of SO and the topical trend in the last decade.

Index Terms— Stack Overflow, question-answer, big data analysis

1. INTRODUCTION

The Question answering sites provide a platform for online users to share and exchange knowledge on a variety of topics, among which SO is specialized for developers and software engineers.

The dataset used in this paper is from Google BigQuery Public Data: Stack Overflow which is updated over time. By the time this paper is completed, available data has ranged from 08-01-2008 to 12-01-2019 with public information of 11 million users and 18 million questions. The data is imported into the HDFS of my GCP cluster and then queried by PySpark.

This paper consists of three parts of analysis, namely, user-based analysis, question-based analysis, and tag-based analysis. In user-based analysis, four types of users are defined, and the user amount of each type is counted. Users within different reputation ranges are expected to behave differently, which would be observed in the amount of the question they ask, the amount of the answer they provide and their posting frequency.

In question-based analysis, the source of the answered and unanswered questions is tracked, which answers the question as "Those answered and unanswered questions are originally raised by which group of users?". Then, the reaction of different group of users toward a new question is explored and the result suggests a pyramid-like interpretation. Meanwhile, some suggestions of how to improve the answer rate in SO are made based on the findings in this part.

In tag-based analysis, the popular taps over the last ten years are selected based on the number of the question related to the tag and the topic trend of these tags is presented. Based on that, variance-based event detection algorithm is applied to show the turning point where users change their focus in SO.

2. USER-BASED ANALYSIS

2.1 User Type

Users can be classified based on their behavior mode in SO. There are four types of user defined in this section, i.e. questioner, answerer, question-and-answerer, and do-nothinger. Questioner is defined to be the user who only asks questions. Answerer refers to the user only answering questions. Question-and-answerer is the user who both ask and answer questions. Do-Nothinger, as the name suggests, never asks or answers any question. Please note that do-nothingers are not really doing nothing in SO, they do not ask or answer question, but they would vote on other questions or answers. The distribution of the four types of users is shown in Table 2.1

Table 2.1 Distribution of four types of users

User Type	Number of Users	Percentage (%)
Questioner	2,277,145	19.19
Answerer	923,351	7.78
Question-and-answerer	1,332,988	11.23
Do-nothinger	7,333,760	61.70
Total	11,867,244	100.00

The result is surprising in terms of the percentage of do-nothinger. There are 61.70% users being inactive all the time never asking or answering any question whereas only 38.30% active users have asked or answered at least one question. One possible explanation of the fact that most users are inactive is that for most users they can always find similar questions in SO, which results in no need for them to post a new question. On the other hand, they are not professional enough to provide answers to other questions, so they are eventually doing nothing.

However, SO is still successful even there are only relatively small part of users being active. The posting frequency of them is expected to be high enough to cancel

the negative impact brought by inactive users. The posting frequency will be discussed in 2.3

2.2 User Reputation

The three most important activities in SO are asking, answering, and editing. A user gains reputation when his post is voted up or the suggested edit is accepted by other users. Basically, the more reputation a user has, the more professional he is.

In this section, every user is labeled as level 1, level 2, level 3, level 4, and level 5 if his reputation is in the range of 1-100, 101-1000, 1001-10000, 10001-10000 or >100000. The level of the user is expected to has an effect on the relative number of the question and answer of that group. The result is shown in Table 2.2

Table 2.2 Effect of reputation on question and answer

Reputation	Users (%)	Question (%)	Answer (%)
1-100	91.94	33.69	9.56
101-1000	6.35	31.35	19.36
1001-10000	1.53	26.86	34.17
100001-100000	0.17	7.70	26.76
>100000	0.0073	0.41	10.15
Total # of each item	11,867,244	18,276,362	27,970,762

Level 1 user is the largest group in SO. They are asking most questions but provide least amount of answers. However, the amount of questions from this group (33.69%) is roughly the same as that of level 2 user (31.35%) whose amount is far fewer than level 1 user. Therefore, it is possible that most level 1 users are do-nothingers. In a word, level 1 users ask the most but answer the least.

There are only 1.7073% professional users (level 3, 4, 5) in SO. The amount of question from these groups are much fewer than that of level 1 and 2 users but they provide more than 70% answers. In a word, professional users ask the least but answer the most.

Level 2 user has almost the same amount of question as level 1 user, but they provide more answers than that of level 1 user. Thus, level 2 user has the characteristic of both level 1 user and professional user. In other words, they ask fairly a lot of questions like level 1 user but also provide many answers like professional user.

2.3 User Posting Frequency

Based on the result from last section, reputation of the user does have an impact on his post. This section is trying to answer the question as "How often do users of each level post?" The result is shown in Table 2.3

Table 2.3 Effect of reputation on posting frequency

Reputation	Post	Per user	Per user per month
1-100	8,831,773	0.81	0.0059
101-1000	11,143,718	14.79	0.11
1001-10000	14,465,667	79.23	0.58
100001-100000	8,892,982	454.02	3.31
>100000	2,912,984	3379.33	24.67
		Avg: 3.65	Avg: 0.027

Although level 1 user is the largest group in SO, they post far less often than any other group. This verifies the assumption that most level 1 users are do-nothingers.

Level 5 user stands out from the table as each user has more than 3370 posts on average, which is 925 times greater than the average, and the posting frequency is 913 times higher than the average. This is good news to SO community because they post very often (almost 1 post / day) to help less professional user with their questions.

Based on the observation of this part, we can conclude that users of SO community are highly polarized in the sense that: 1) In terms of user type, there are 38% doers but 62% do-nothingers; 2) More than 90% users are novice user while there are only 1.7% professional users; 3) Tiny part of users (level 5 users) are posting much frequently than other users.

3. QUESTION-BASED ANALYSIS

Question is no doubt one of the most important components of SO. 86% questions on SO are answered in the past 11 years, which is relatively good, but the answer rate of each year is decreasing as shown in Fig 3.1

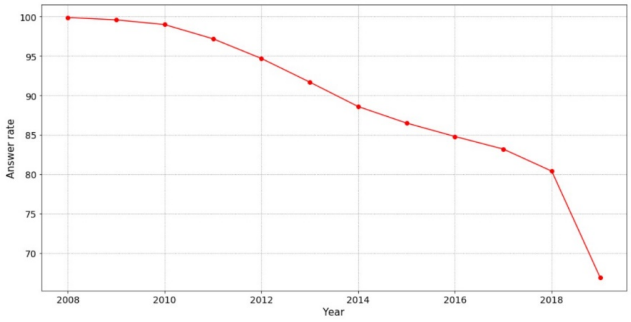

Fig 3.1 Answer rate from 2008 to 2019.

This part of work is trying to give advice of how to improve the answer rate based on the analyze of answered and unanswered questions.

3.1 Answered and Unanswered Question

It is helpful to know the source of answered and unanswered question or, in other words, "Where are the answered and unanswered question from?"

The relative amount of answered and unanswered question from each level of user is shown in Table 3.1

Table 3.1 Effect of reputation on answered and unanswered question

Reputation	Question (%)	Answered (%)	Unanswered (%)	Answer rate (%)
1-100	33.69	9.56	46.70	80.72
101-1000	31.35	19.36	30.74	86.36
1001-10000	26.86	34.17	19.18	90.07
100001-100000	7.70	26.76	3.28	94.07
>100000	0.41	10.15	0.10	96.61
	Questions: 18,276,362	Answered questions: 27,970,762	Unanswered questions: 2,542,464	Avg: 86.09

The result shows that 60% answered questions are from level 3 and 4 users and 76% unanswered questions are from level 1 and 2 users.

level 1 user has most question (33.69%), but least amount of answered questions (9.56%) are originally raised by them. Besides, almost half of the unanswered questions are raised by level 1 user.

Level 2 user has similar amount of questions as level 1 user, but the relative amount of answered question of them is 10% higher than that of level user and the unanswered questions from them is 16% less. One could assume that level 2 user are better at asking question and getting themselves understood.

In fact, it is noticeable that questions from higher-level user are more likely to be answered (as answer rate increases from level 1 user to level 5 user). This is reasonable as higher-level user is supposed to be asking a clear question, which increase the possibility of the question to be answered.

Although there are already many guidelines of how to ask a good question in SO, they do not seem to work well since level 1 and 2 users are still the main source of unanswered questions. Thus, to improve the answer rate, SO could make special guideline exclusively for those novice users.

3.2 SO Question-answering Pyramid

This section mainly focusses on what will happen when a new question is asked. Some of the questions will never be answered but for those answered question, they will receive the first answer within 24 hours on average. The answered questions with up to ten answers are discussed in this section.

To answer a general question as "What will happen after a new question is asked?", it is safe to break down the question into multiple sub-questions, namely, "What is the reputation of the user providing the 1st answer?", "What is the reputation of the user providing the 2nd answer?" … "What is the reputation of the user providing the 10th answer?", from which the reaction of users from every level can be viewed, which further indicates how users in SO community react to a new question.

The absolute number of answer-provider in the answer of each order is shown in Appendix Fig A1. Level 1 and 2 users are relatively increasing in later-order answer, level 3 users are the main answer-provider in the answer of any order, level 4 and 5 users are relatively decreasing in later-order answer.

The same result is replotted to show the relative number of answer-provider from each level in Fig 3.2

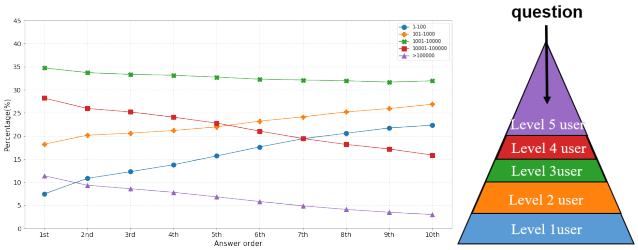

Fig 3.2 User percentage in the answer of each order and the pyramid structure of users' reaction to a new question

The relative number of level 3 users, the green line, remain unchanged while higher level user like the purple and the red lines are decreasing. Lower level user like the blue line is increasing. The dynamic of users' reaction to a new question is like a pyramid. Higher level user will provide the first several answers which are then followed by lower level users who try to complete or improve previous answerers. However, this does not suggest that SO is working explicitly like this pyramid. It just suggests one of the possible interpretations of how SO users will react to a new question.

Based on this, to improve answer rate, SO administrator could try posting those unanswered questions to higher level user to speed up the reaction of the whole community toward a new question.

4. TAG-BASED ANALYSIS

When asking a question, the questioner can choose up to 5 tags to label the question. A tag is a word or phrase that describes the topic of the question, which is a mean of connecting experts with questions they will be able to answer by sorting questions into specific, well-defined categories. Thus, tag can be used as an indicator of what developers care about.

4.1 Topical Trend

There are more than 50,000 tags in SO. Some tags like javascript are very popular among developer while the other receive much less attention. The number of questions related to the top 25, 50 and 100 tags are shown in Table 4.1

Table 4.1 The amount of questions related to the top N tags

Top N tag	Relevant Questions	Percentage (%)
25	11,736,633	63.11
50	13,132,132	70.61
100	14,178,585	76.24

Based on the table above, only the top 25 tags are discussed in this part as they covered more than 60% questions and increasing the number of tags does not make much difference.

To reflect the topical trend over the 11 years, a sliding time window schema shown in Fig 4.1 is used.

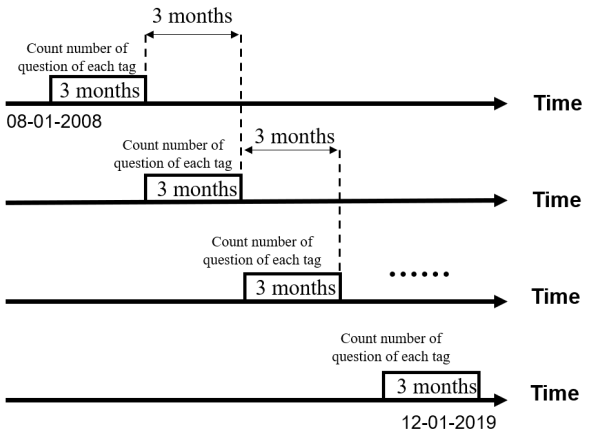

Fig 4.1 The time window schema.

The window is fixed to be three-month long where the number of questions of each tag is counted. The window starts from 08-01-2008 and move 3 months forward at a time until it reaches 12-01-2019. The topical trend obtained by this method is shown in Fig 4.2

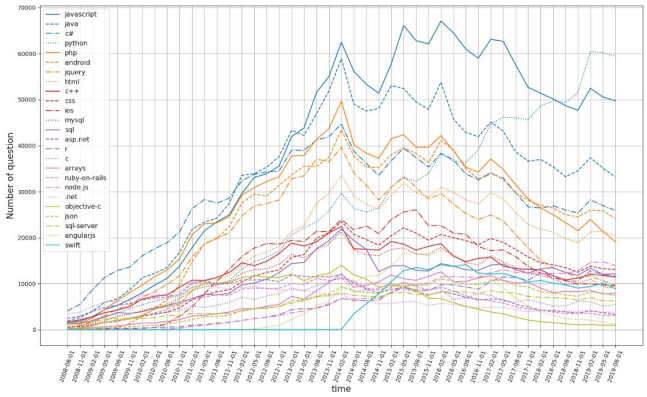

Fig 4.2 Absolute number of questions of top 25 tags

Based on the graph above, javascript used to be the most popular tag. Python has been receiving more and more attention over the years and surpassed javascript on 08-01-2018, becoming the hottest topic.

Besides, Php curve, c# curve and android curve share similar shape. This indicates that the three tags are often used to label a same question, which is reasonable as it is very possible to run into a problem associated with php, c# and android at the same time when developing an android-based web application.

One should be aware that the growing number of questions is the result of two factors: one is the growing number of users and the other is the tag itself becoming more and more popular. In order to cancel the influence brought by growing number of users and focus on the tag itself, the same result is normalized by the total number of questions within each 3 months and shown in Fig 4.3

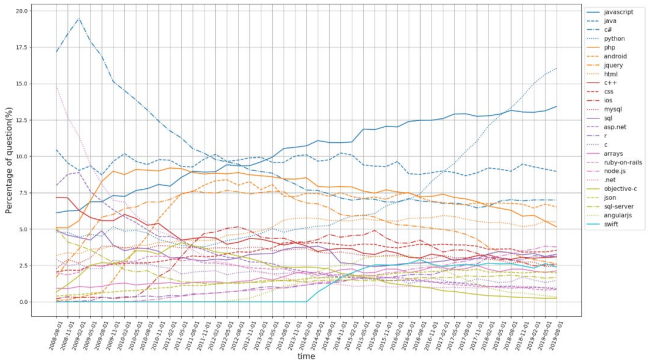

Fig 4.3 Relative number of questions of top 25 tags

Now it is clear that c# used to be even more popular than any nowadays popular language, which is not easy to be noticed before normalization. However, developers' focus is gradually changing from c# to javascript and python possibly because dynamic language like javacript and python are easier to learn and use.

Java curve is steady over the years with its relative number remaining around 10%. This suggests that java is always an important topic for developers. However, c is not the same fortunate as java, which has been decreasing from 15% to 1.7%.

4.2 Variance-based Event Detection

Based on the result in the last section, some nowadays popular tags were not the same popular in the past. That means some events must have happened where those tags gain their popularity. This is the motivation of this section, to detect the event that makes a change.

If some event happens, for example python 3.0 is released, the amount of python questions is expected to be much more than before as python 3.0 is not compatible with its older version. After some time, users have run out of the question related to this event so that the amount of python question should return to normal level. In other words, the variance of the amount of the question is supposed to be higher than usual when an event happens. This is the key assumption of the variance-based event detection algorithm employed in this section. The mechanism of it is shown in Fig 4.4

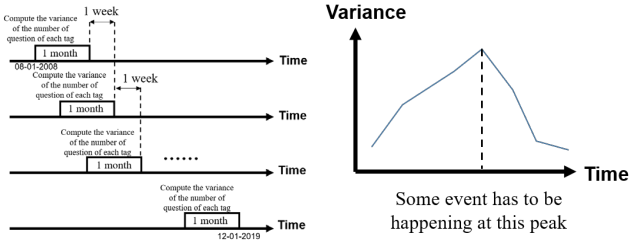

Fig 4.4 Variance-based event detection mechanism

It is similar with the sliding time window schema as in Fig 4.1. The time window is 1 month long where instead of counting the amount of questions of each tag, the variance of the amount of question is computed. The time window is moving forward 1 week at a time with 2 weeks overlapping with the previous window. The event is supposed to be detected at the time when there is a variance peak. The result is shown in Fig 4.5

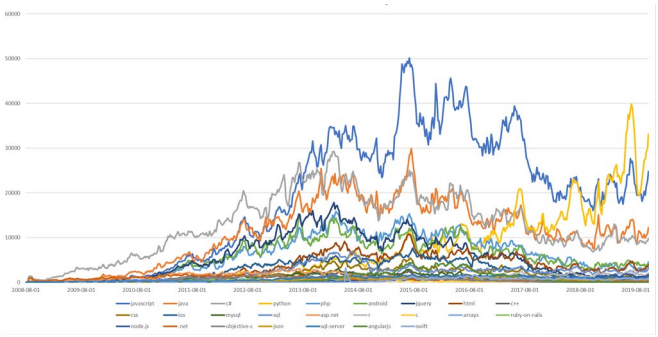

Fig 4.5 Tag variance trend

For javascript, the strongest peak is around the May and June in 2015 probably because at that time an important framework for javascript called Polymer is released by google, and in the following month ECMAScript 6 is also released, which is a standard for scripting language like javascript.

Python curve has many prominent peaks which could be the turning point where python receives more attention. However, it is difficult to find out what events cause the variance peaks. There are thousands of hundreds of questions within each time window, which is infeasible for me alone to summarize the questions based on their semantic.

This event-detecting algorithm is simple and intuitive, but it just tells us there should be an event or some events. In fact, it is very difficult to find out which event(s) lead to the variance peak because on the one hand, it is tedious to manually summarize the question based on their semantic and developing a program to automate this process would be extremely challenging. On the other hand, we do not know beforehand the variance is the result of a single strong event or multiple weak events. To interpret the result requires the knowledge out of big data technique, so this paper just ends up here without any further exploration.

5. CONCLUSION

This paper analyzed SO community from three aspects: user, question and tag. SO users are highly polarized in terms of type, reputation and posting frequency. Many users are inactive with their reputation being in the range of 1-100 and post much less often while only a relatively small group of users are active with the reputation greater than 1000 and post very often.

Most unanswered questions are originally raised by novice user, which suggest that those users are lack of experience of how to ask a question nicely. The relative number of the answer-provider in the answer of each order suggests an implicit question-answering pyramid in SO based on which, expose the unanswered question to more professional users is potentially a way to improve answer rate.

Finally, SO has many tags but only the top 25 of them are popular. The topical trend reflects developers' most concerned topic over time and the variance-based event detection suggests the turning point where developers moved their attention.

6. REFERENCES

[1] Asaduzzaman, M., Mashiyat, A.S., Roy, C.K. and Schneider, K.A., 2013, May. Answering questions about unanswered questions of stack overflow. In 2013 10th Working Conference on Mining Software Repositories (MSR) (pp. 97-100). IEEE.

[2] Anderson, A., Huttenlocher, D., Kleinberg, J. and Leskovec, J., 2012, August. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 850-858).

[3] Cheng, J., Danescu-Niculescu-Mizil, C. and Leskovec, J., 2014, May. How community feedback shapes user behavior. In Eighth International AAAI Conference on Weblogs and Social Media.

[4] Wang, S., Lo, D. and Jiang, L., 2013, March. An empirical study on developer interactions in StackOverflow. In Proceedings of the 28th Annual ACM Symposium on Applied Computing (pp. 1019-1024).

[5] Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P. and Faloutsos, C., 2013, August. Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013) (pp. 886-893). IEEE.

[6] Barua, A., Thomas, S.W. and Hassan, A.E., 2014. What are developers talking about? an analysis of topics and trends in stack overflow. Empirical Software Engineering, 19(3), pp.619-654.

Appendix

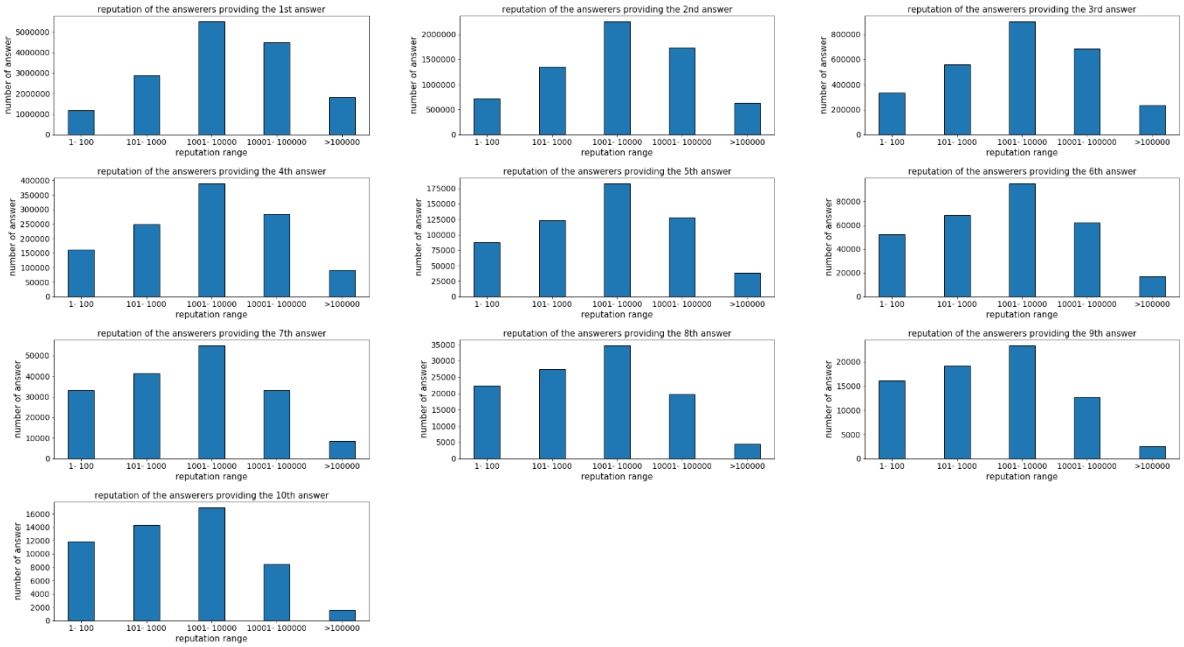

Fig. A1. Reputation of the answer provider in the answer of each order.