ECS 260 Software Engineering

Final Project : Interactions among Experienced Users in StackOverflow

Shu-Wei Hsu, Yun-Chieh Sung

Background

Stack Overflow is a Stack Exchange Network featuring various topics in computer programming founded by Jeff Atwood and Joel Spolsky in 2008. Stack Overflow now has over 1 million registered users and more than 3.5 million questions, and with our given data, more than 5 million posts (including questions and answers). Through membership and with consents to Creative Commons License, registered users are free to ask, answer, and vote questions with feedbacks from other users. Stack Overflow designs a mechanism of reputation points and badges. Users are awarded with reputation points if they give an answer to someone's question with upvotes or get their answer accepted. They may receive badges regarding to their contributions and have additional privileges. The primary way to gain reputation is as listed below:

Answer is voted up	+10	
Question is voted up	+5	
Answer is accepted	+15	(+2 to acceptor)
Question is voted down	-2	
Answer is voted down	-2	(-1 to voter)

Figure 1. The Mechanism of Gaining Reputation in Stack Overflow

In a word, as a concept of contributions with the mechanism of reputation points and badges, it is the assumption that users who answered more questions and got accepted have more computer programming skills, experienced, and deserve higher reputation. Moreover, people with high reputation are supposed to answer more questions or give a questions with higher view counts, which means a more useful information to the public, and therefore, a valuable contribution. However, is it always true?

Research Questions

According to the design of Stack Overflow, personal reputation is relative to contribution and is composited by user voting and accepting the answers. There are several questions drawn by the reputation mechanism whether the users with high reputation deserve the status and privileges due to their contributions:

1. Are the users with higher reputations usually also with higher view counts? which means a more useful information to the public, and vice versa?

2. Are the answers provided by users with higher reputations, who are considered more experienced, would get accepted more easily (accepted rate), and vice versa?

3. Do registered users with higher reputations interact with each other more often? In addition, the user who interacts with the other user more can usually answer with higher score (higher upvote count and less downvote count) in the questions the other user asked?

Q: I need to match all of these...	
id	1732348
parentid	[Null]
acceptedanswerid	1732454
score	1330
viewcount	525534
owneruserid	142233
answercount	36
displayname	***Jeff***
reputation	***984***

A: You can't parse [X]HTML with...	
id	1732454
parentid	1732348
acceptedanswerid	0
score	4432
viewcount	0
owneruserid	18936
answercount	0
displayname	***bobince***
reputation	***124849***

Figure 2. Two Posts of Question and Relevant Accepted Answer of The Title "RegEx match open tags except XHTML self-contained tags" from "posts" and "users" (*bold italic*) Tables

Experimental Method

1. As a post of question shows on the left in Figure 2 (Unused or less related data are left out), we query the users with top reputation from "users" tables and generate the other users higher ranked in sum of view count(SOV) from "posts" table. The SOV is referred to the sum of viewcount, which is an attribute of one's post. By overlapping the two group of users, we produce a list of key users and we expect this list to cover the majority of the users of high reputation and users of high SOV. Which indicates that the users with top reputation usually give more useful information to the public.

2. When an answer of that post of question (as showed on the right in Figure 2, the attribute "parentid" is set with the id of the original question) has been accepted, the id of the answer is stored to the attribute "acceptedanswerid" of the original question. Therefore, in the same manner of retrieving the posts of questions from users with top reputations, we retrieve the post of answers and

map to the original questions by the attribute “parentid” to see if the answers are also accepted. We use the results to calculate accepted rate and to find the relationship between these two factors.

3. Furthermore, we look deep into the interaction within registered users with higher reputation. We pick a group of high reputation people as sample set. As shown in Figure 3, If one answer the other’s question, we use an arrow to represent the action. One user who give an answer to the other, may also receive an answer from the same person in the past or in the future due to their both being skilled in the same area because of high reputations (the upper two in Figure 3). We expect most of users will interact with others bi-directly rather then just interacting in single way (the lower two in Figure 3).

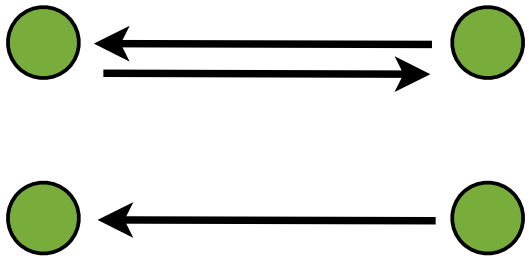

Figure 3. Interactions Between Four High Reputation Users (Giving Answers)

Finding

With containing millions of users, StackOverflow has 1499998 posts and 3669188 replies. The following is the ratio of total posts and replies (Figure 4). We can figure that the majority of articles are referred to reply questions.

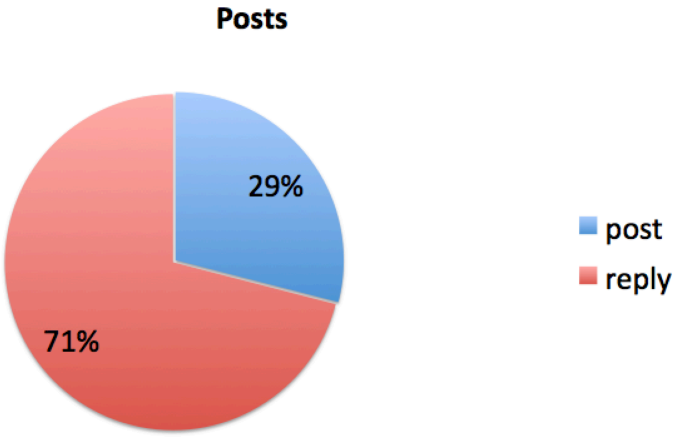

Figure 4. Interactions Between Four High Reputation Users (Giving Answers)

In the first research question, we queried the data from “posts” and “users” and generate the top 10000 users with highest reputations and top 10000 users with their post of highest sum of view count. We found out that in most of case, the user with lower reputation would have higher sum of viewcount. Which can be interpreted as to post more general question.

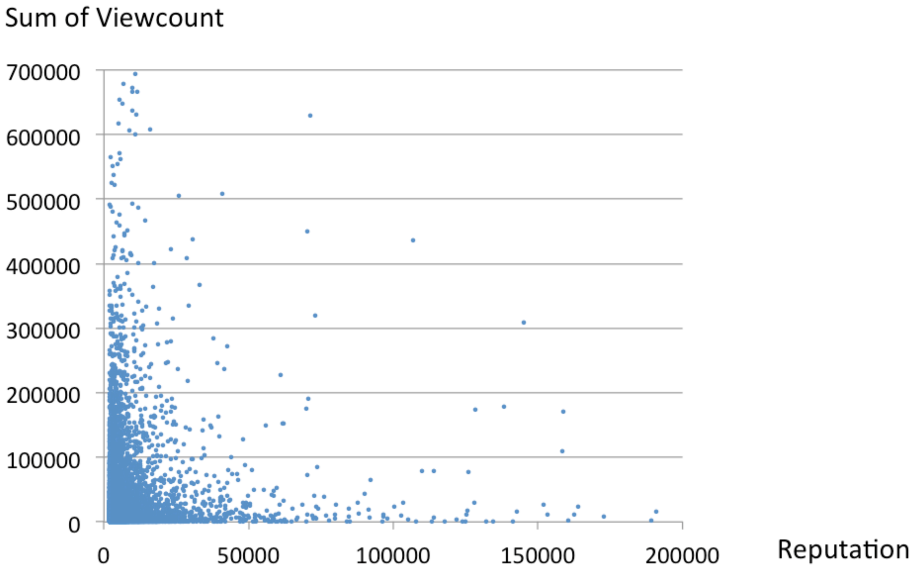

Figure 5.1 Users in top 10000 reputations and their relative sum of view counts in their post

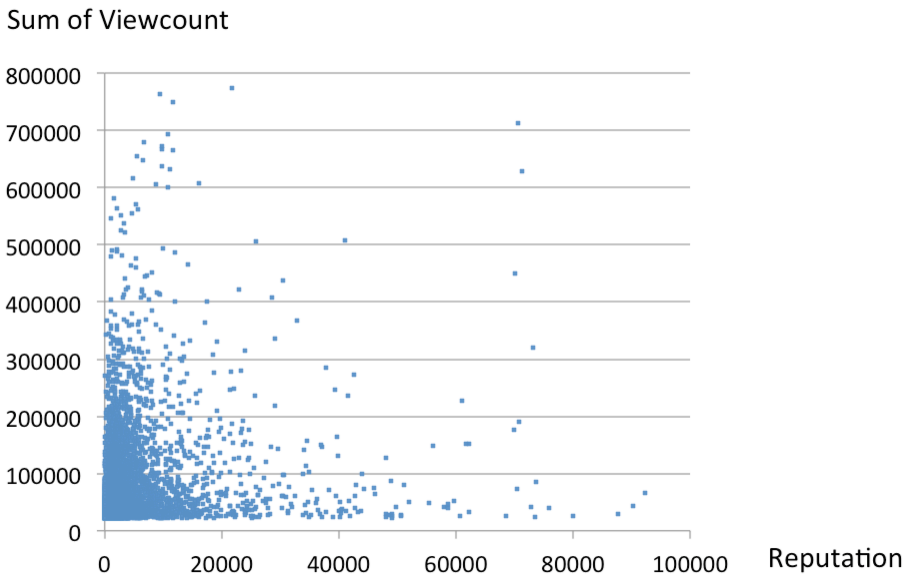

Figure 5.2 Users in top 10000 sum of view counts in their post and relative reputations

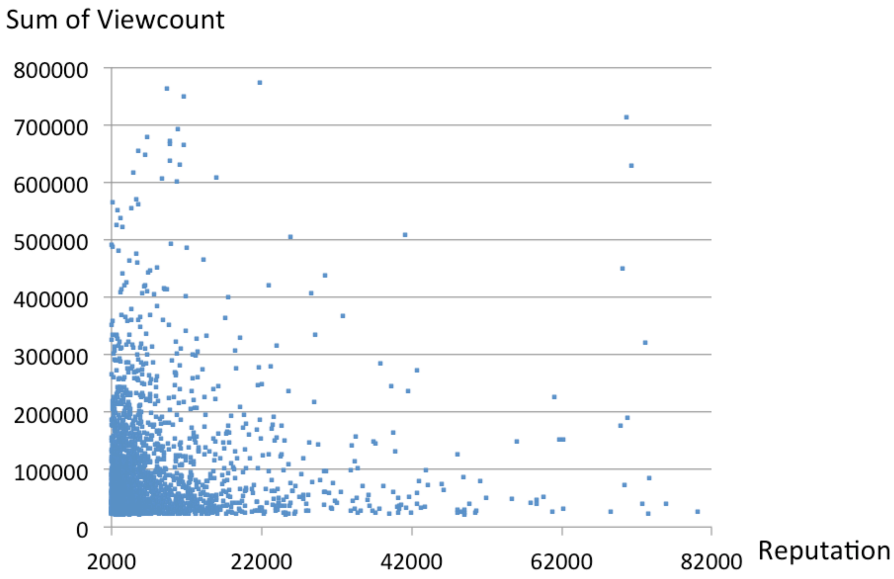

Figure 5.3 Users both in top 10000 sum of view counts in their post and top 10000 reputations

In the second research question, we employed users in top 10000 and top 100000 reputations to identify the ratio of answer replied by user in higher reputation to be judged as the correct answer. The results shows that the answer replied by user who has the highest reputation usually is the best answer in the discussion (accepted answer).

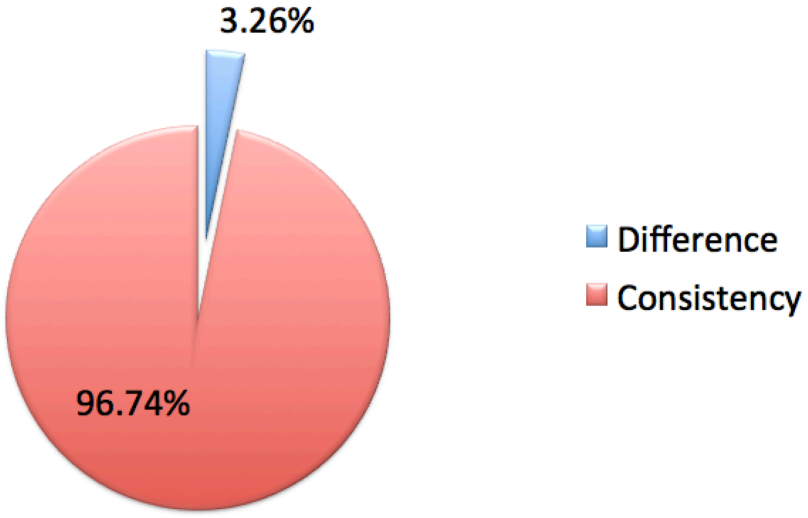

Figure 6.1 In the rate of 96.74%, the reply by user in highest reputation (in top 10000 reputations) can be the best answer in discussion (accepted answer)

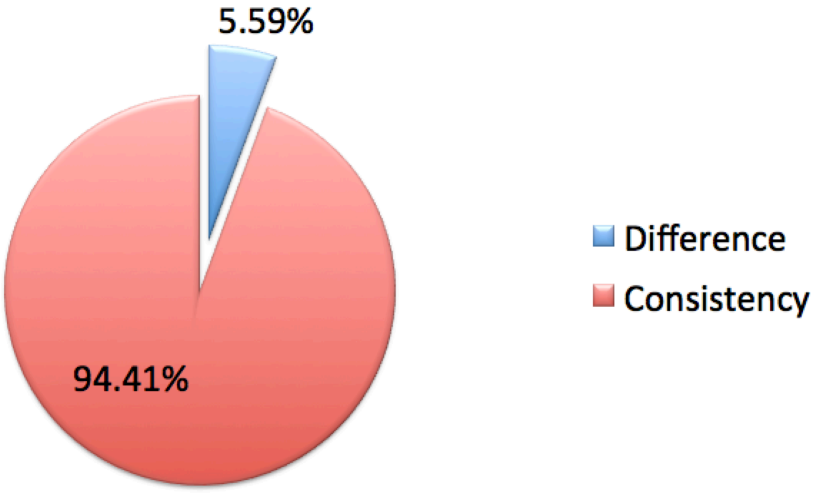

Figure 6.2 In the rate of 94.41%, the reply by user in highest reputation (in top 100000 reputations) can be the best answer in discussion (accepted answer).

In the less research question, we found out that reputation has almost nothing to do with user id creation date. Furthermore, we employed the open source, D3.js, to visualize the interaction between experienced users, who is in higher reputation. As In Figure 7.1 and Figure 7.2, we concluded that all of interactions are one direction interaction and most of interactions are distributed.

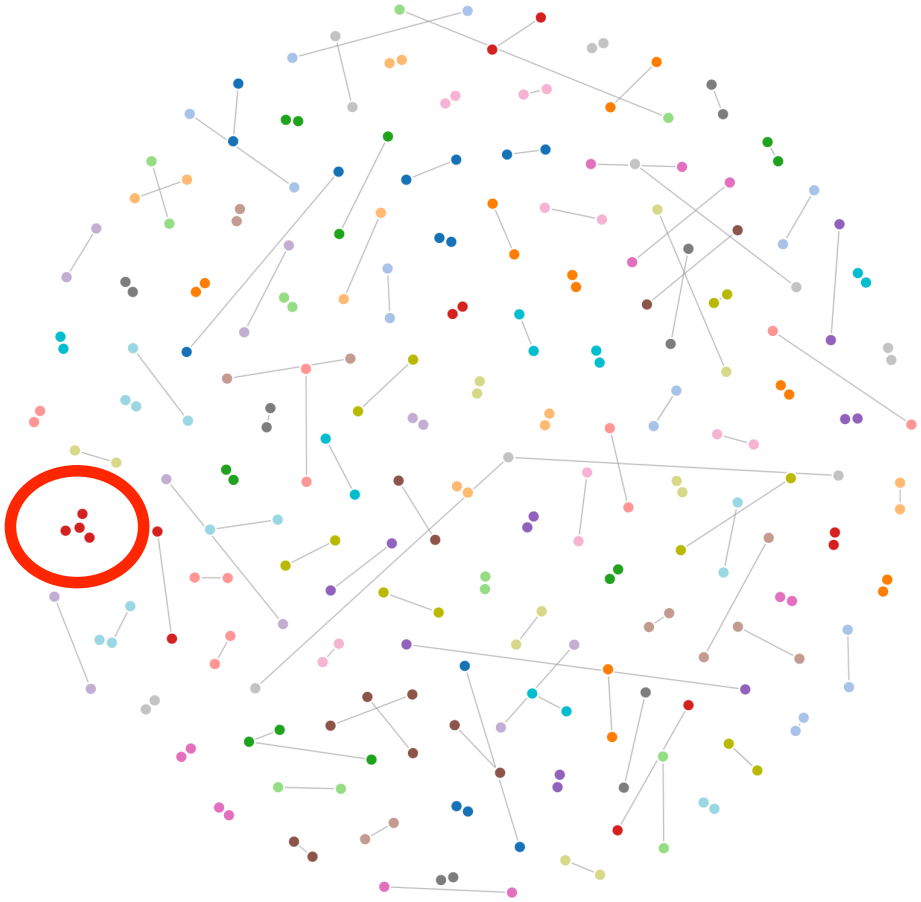

Figure 7.1 Interaction between users who are in top 1000 reputations.

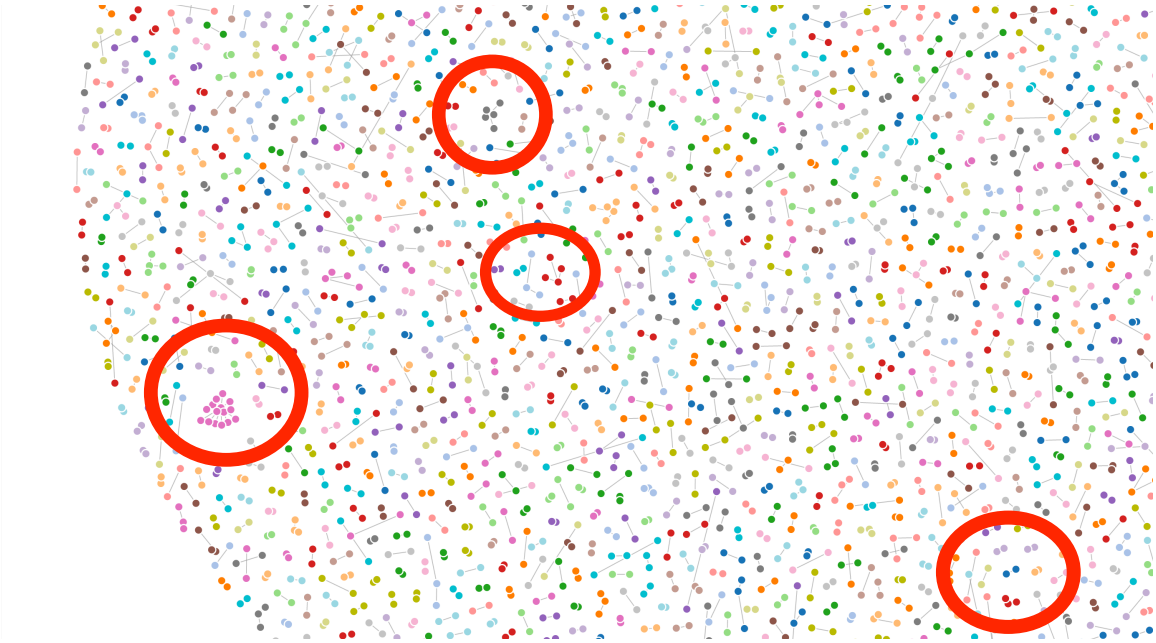

Figure 7.2 Interaction between users who are in top 10000 reputations.

Conclusion

We design a detail study of the attributes of “users” and “posts” table. A focus is on the reputation of user and evaluation of the contribution of Stack Overflow’s reputation points criteria, through the inspections of view count of posted questions and answer-accepted rate. Furthermore, we do more research in interaction among high reputation users. Through our approach, we want to generate a user set who always give positive answers and questions, and contribute more valuable information to public then others in Stack Overflow. In addition we conclude that: the user with lower reputation would post more general question (identified by view count); reputation has almost nothing to do with user id creation date; the answer with highest reputation has high possibility to be the accept answer; users of higher reputation actually post more questions (less replies); there is less dual interaction among experience users (only one direction).

Future Work

So far, we can only find out less interaction between high reputation users, which has the question-and-answer activities more distributed and independent than our assumption. Therefore, we are trying to do more research in the following aspects in the future:

1. Comparison of the reputations difference between the post and reply.
2. Group the posts with categories (tags)
3. View counts versus categories
4. Weight the interactions by the key words in titles and contexts (title, body)