

# Is it better to build multi-language Stackoverflow site?

Australian National University  
u5870997@anu.edu.com

June 8, 2017

Abstract

## I. INTRODUCTION

## II. CONTENTS

- Link
- User
- Post
- Comment
- Tag

## III. LINKS

This part of work is based on the links of stackoverflow and Russian stackoverflow. According to the previous results by some other similar research, links are good proof to present the user activity and community activity. Combine the amount of links and the activities of link between these two communities can provide some evidence for the necessity of a multi-language sub site. First of all, we need to find out how many posts and comments used links, and how many people links there evidence in their answers and questions. Then the analysis would be presented in some charts to show the results.

Currently, there are 34,857,917 posts and 55,852,373 comments in Stackoverflow site while 280,424 posts and 499,186 comments in Russian Stackoverflow subsite. Basically, they are divided into two sets, one is the set of posts and comments who use links to stackoverflow.com and the other is those to ru.stackoverflow.com. Some statistics are presented in figures below.

For the posts and comments in Stackoverflow:

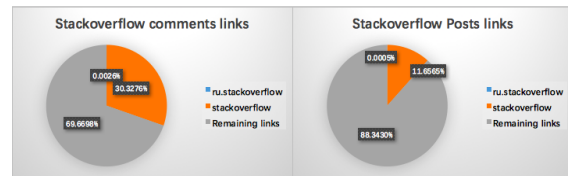
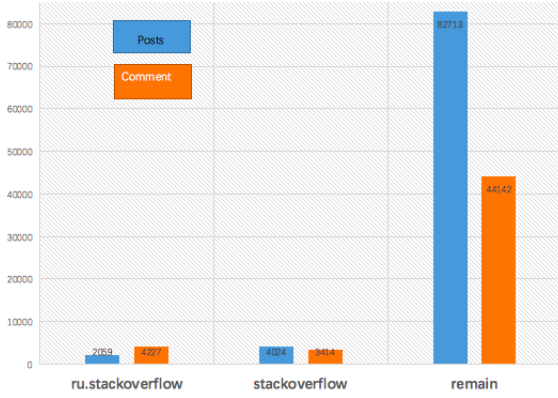
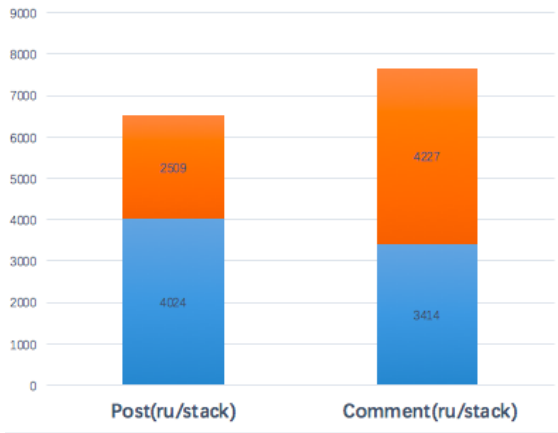


Figure 1: Links in Stackoverflow

Similarly, for Russian Stackoverflow:



**Figure 2:** Amount of links actively in Russian Stackoverflow



**Figure 3:** Compare between the amount of links to stackoverflow site and Russian subsite

Apparently, in stackoverflow set, 12,220,205 posts used links in their body text. 11.66% of them link to stackoverflow itself, while only 67 links are lead to Russian subsite. In Russian stackoverflow, 88,796 posts used links in their body text. 4.53% of them link to stackoverflow main site, while 2.32% of them link to Russian subsite itself. The result of the link activities shows that Russian Stackoverflow subsite is relatively independent community, which support itself by a lot of link activities instead of making the majority of their references to the main Stackoverflow site. However, in some ways the Russian Stackoverflow subsite has a non-negligible demand of knowledge reference to the main site according to the compare of destination between the amount of links.

#### IV. USERS

It has been widely believed that users data is highly important for the community analysis. User activity can be revealed by data, which have already been structured into different types. Hence, the analysis of user would be

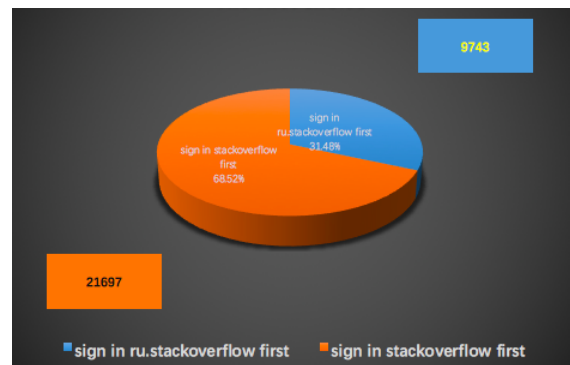
expanded with different aspects.

First of all, the intersection of these two sets is the foundation of this part of research, which is the users that owns the main site account and Russian subsite account at the same time. A point worth mentioning is that not only people from Russia use Russian for introduction and answers, but also some other countries like Ukraine and Belarus. That is the user base called "Russian Speakers".

Stackoverflow main site has 6,833,276 users, while Ru.stackoverflow has 65,623 users at the same time. Each stackexchange user account has a unique accountId, which is always the same if the user sign up for a subsite account like Russian or Spanish Stackoverflow. Using this information, the number of the intersection can be found as 35,751, which is nearly 54.48% of the total number of users in Russian Stackoverflow.

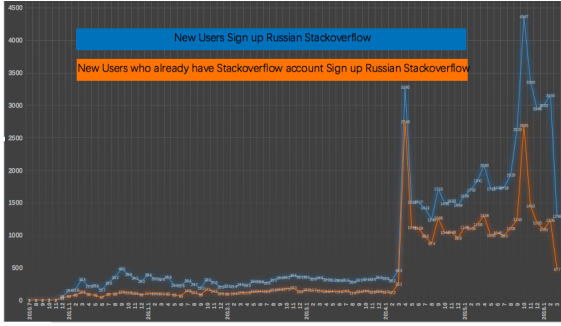
##### i. CreationDate

Focusing on this intersection user base, the user movement is calculated by their account creation date of each site. As the result shows in figure 4, 11,253 people have their Russian Stackoverflow account first and then sign up for a Stackoverflow account. While the other 24,498 users are the opposite of them. On the one hand, user base is segregated by the new subsite and the number of users leaving the main site to their native language subsite is growing with the time goes by. On the other hand, when 3 out of 10 users flow to Stackoverflow from Russian Stackoverflow, the other 7 of these 10 people flow to Russian stackoverflow from Stackoverflow, which means Russian site surprisingly offers a lot of users as feedback to Stackoverflow.



**Figure 4:** Creation date compare

Here is the statistics for the users sign up Russian Stackoverflow in figure 5.



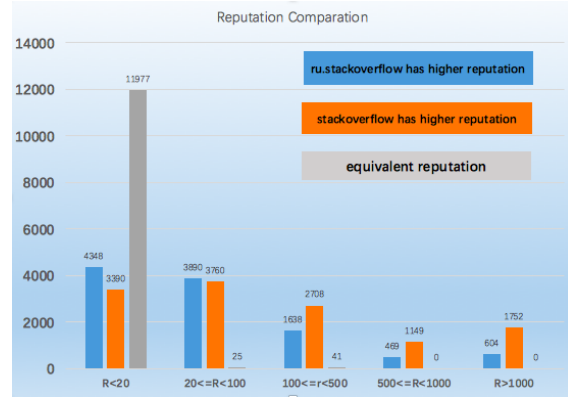
**Figure 5:** User movement for signing up Russian Stackoverflow in each month

## ii. Location and AboutMe

Basically, all the users who can speak Russian can be found in a simple way. The Russian Alphabet can filter all the people using Russian in their Displayname or self introduction of the main Stackoverflow account. In this way, 19,300 users can be seen as a set. Surprisingly, only 3,946 of them have a Russian stackoverflow account, which means 79.55% of these Russian speakers choose to use Stackoverflow site only. There are not a pretty huge number of users use their native language, which is Russian, in Stackoverflow site. And compare this number with the total number of Russian Stackoverflow user base, it is still not a big number. The conclusion is that the majority of Russian Speakers prefer to use English as the only general language on the main site, and the fragmenting level is slight.

## iii. Reputation

Reputation is a very valuable information that can be used to reveal the active level of users. The assumption is the user whose reputation is under 20 will be seen as an inactive user. As the result shows in figure 6, a huge number of inactive users, which includes a number of barely used accounts whose reputation is 1.

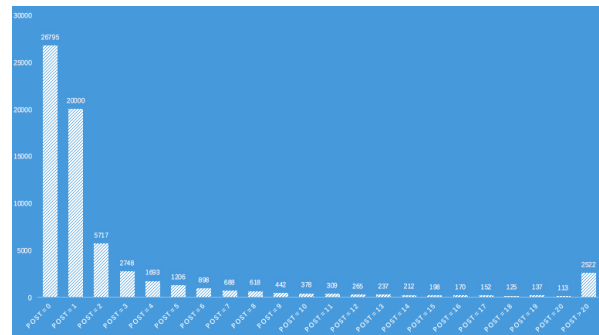


**Figure 6:** Reputation compare

And the remaining data shows there are more users' stackoverflow account owns a higher reputation than their Russian ones. Only 8,360 users in the 31,440 overlapping set have a over 100 reputation in either of their accounts, and in those 24,498 users who migrated from Stackoverflow to Russian Stackoverflow 6,489 users have a over 100 reputation, which means they are active users.

## V. POSTS

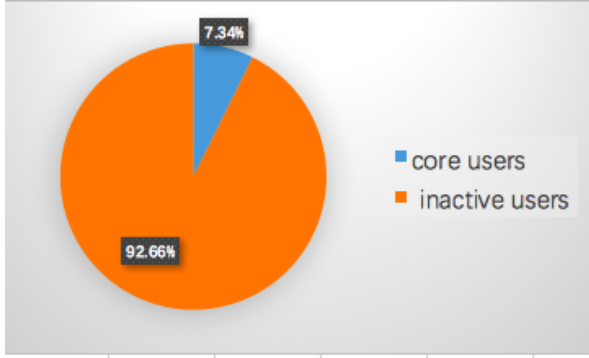
Generally, for each single community, core user groups are the foundation because of their contribution like a large number of posts and comments. Similarly, in figure 7, for the 65623 overlapping users, we can simply get a group of core user by calculating their posts. For the Russian users, simply calculating their post count and the results are presented in figure 8. This result is very consistent with the twenty-eight law, that is, about twenty percent of the user made a contribution of about eighty percent



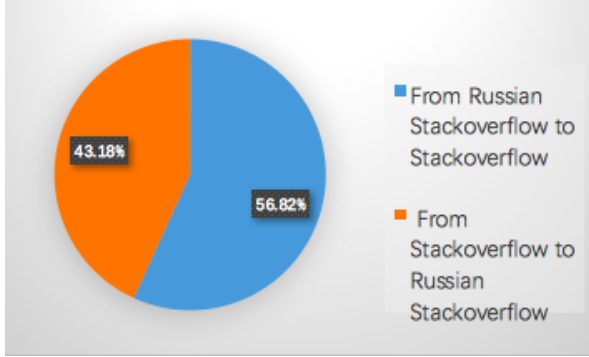
**Figure 7:** russian users posts count result

So we can see that 4818 users who contributed more than 10(can equal) posts own 237,557 posts, which is almost 76.8% of the total amount of Russian posts. So it is better to make an assumption that these kind of users are the core

users for the Russian sub site. On the other hand, we need to use their creation date again to determine their direction of migration between these two sites, which is showed in figure 9.



**Figure 8:** Core users that owns both Stackoverflow account and Russian Stackoverflow account at the same time



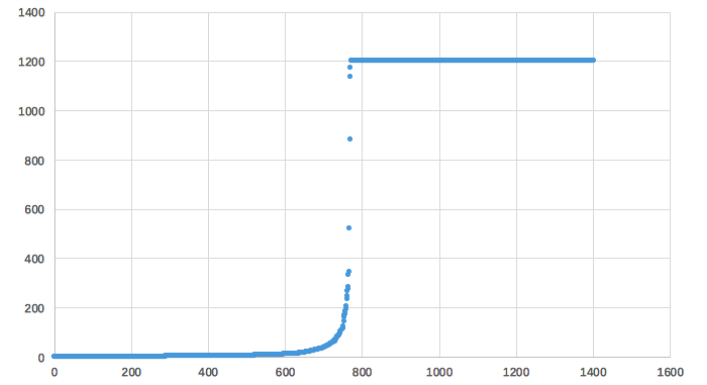
**Figure 9:** The compare of migration direction

As the results above, 1845 out of 11253 overlap users who migrated to Stackoverflow, which is 16.40% are proved still be core users in Russian stackoverflow site, while 1402 out of overlap 24498 users who migrated from Stackoverflow, which is 5.72% are proved to be core users. There are 310,539 posts in Russian Stackoverflow, and 237,557 posts in Russian Stackoverflow are contributed by the core user group that owns 4818 people. In other words, the 7.34% core users contribute 76.50% of the total posts in Russian Stackoverflow.

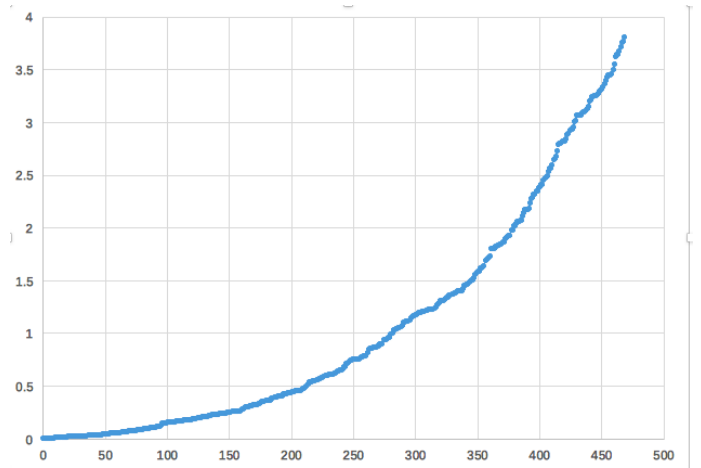
For the 1402 core users who already have Stackoverflow account before their creation of Russian Stackoverflow account, the compare between the post activities before their migration with the post activities after their migration can definitely reveal their emphasized aspect. The idea is that the object of reference of the users is themselves. To compare the post frequency before and after they create a new subsite account, which called user migration.

$$f(x) = \frac{\text{PostCountAfterMigration} / \text{TimeLengthAfterMigration}}{\text{PostCountBeforeMigration} / \text{TimeLengthBeforeMigration}}$$

Briefly, for all the Russian users, the result of their active proportion is in figure 10. We can see no matter which domain we choose, it cannot change the trend of the user active ratio. The main trend is linear increasing.



(a) A general figure for all users

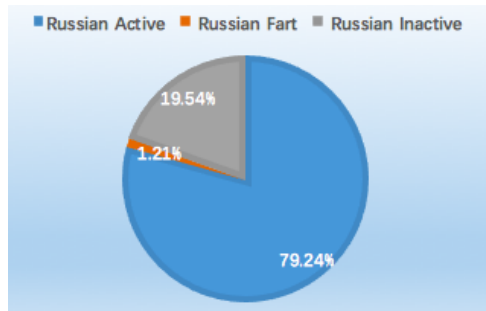


(b) Active ratio close to 1

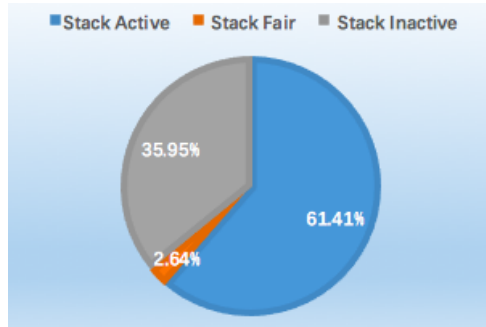
**Figure 10:** active ratio chart

## i. PostCount

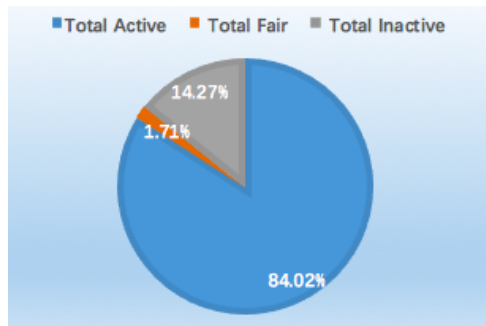
As the result in figure 10, we can easily determine that 0.9 1.1 is the domain of fair activity area. So there are three level of activity, the first one is active, which means the count of users' post is above 1.1 multiple of their previous one. The second one is fair, which is between 0.9 and 1.1. The third one is inactive, which is under 0.9. The result is in figure 11.



(a) Russian Post Activity.

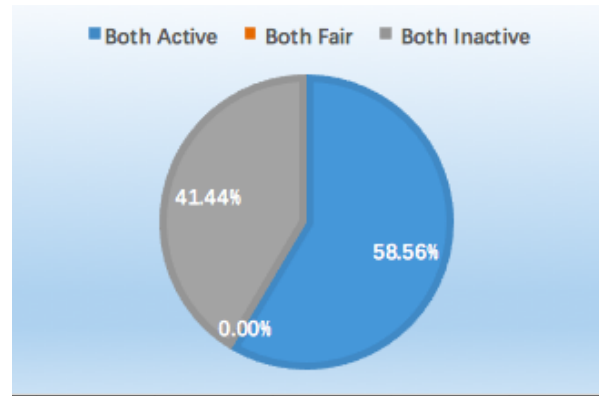


(b) StackOverflow Post Activity.



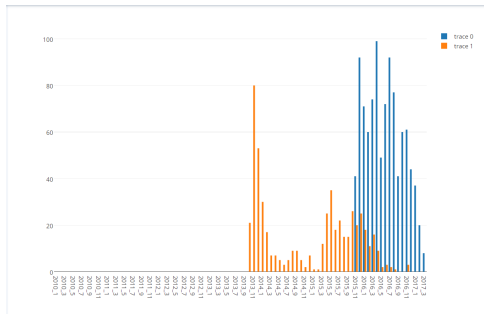
(c) Total Post Activity.

**Figure 11: User Post Activity Result.**

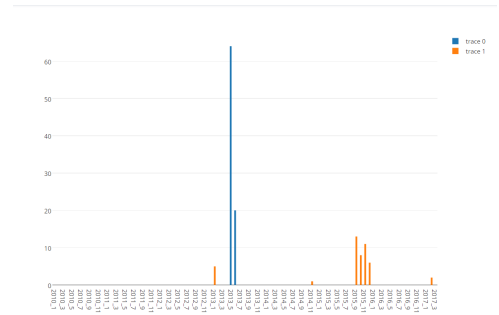


**Figure 12: Overall result**

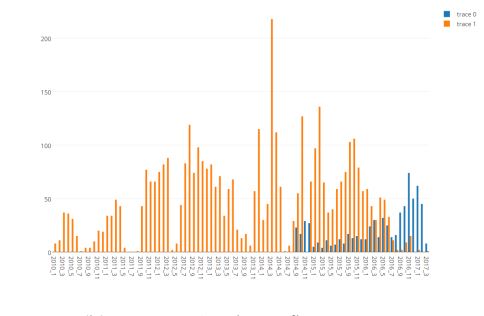
And the in these 1402 users who own both Stackoverflow account and Russian Stackoverflow account and migrated from Stackoverflow site to Russian sub site. 861 of them are both on both sites after their migration, and 3 of them are fair, while 200 of them are inactive. The result is in figure 12.



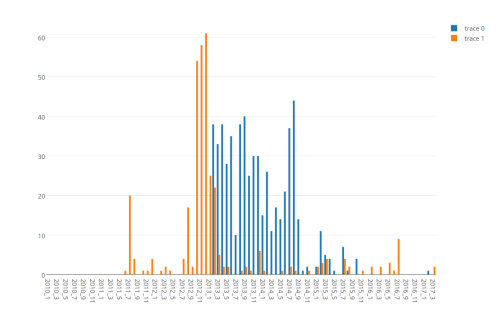
(a) Russian Stackoverflow Active.



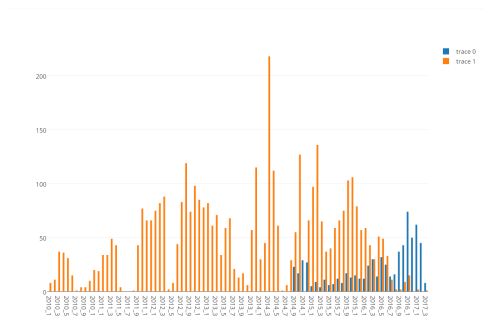
(a) Total Active.



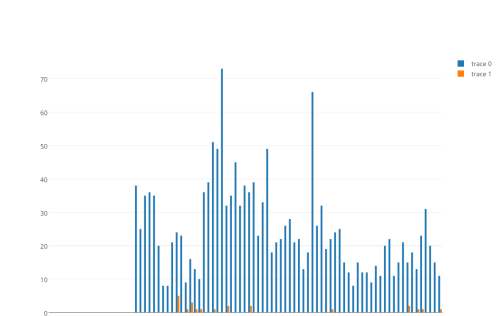
(b) Russian Stackoverflow Inactive.



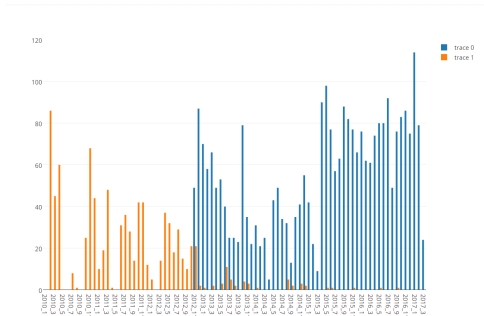
(b) Total Inactive.



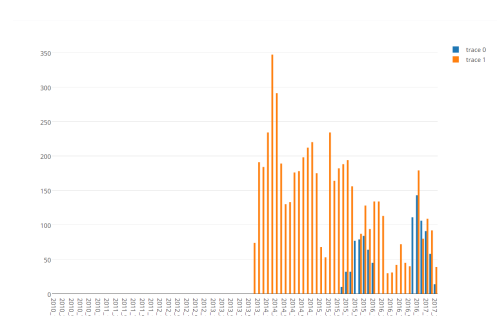
(c) Stackoverflow Active.



(c) Both Active.



(d) Stackoverflow Inactive.

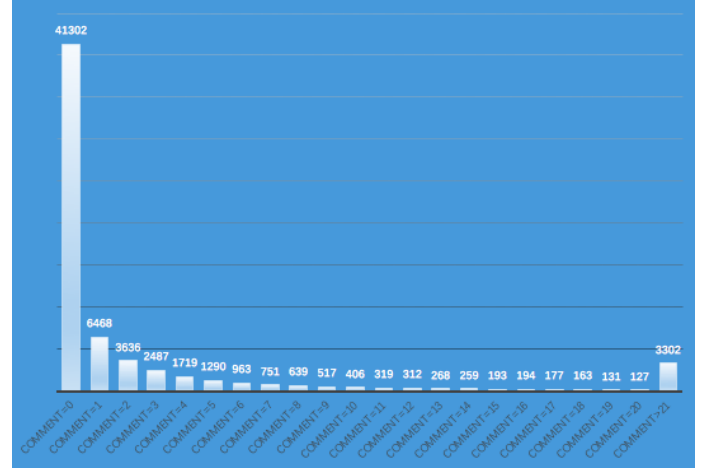


(d) Both Inactive.

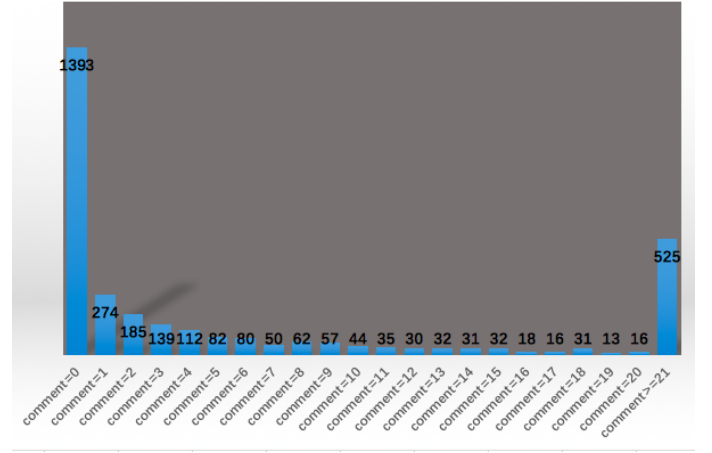
Figure 13: 8 examples(Part 1)

Figure 14: 8 examples(Part 2)

## VI. COMMENTS



(a) Comment amount in Russian Stackoverflow



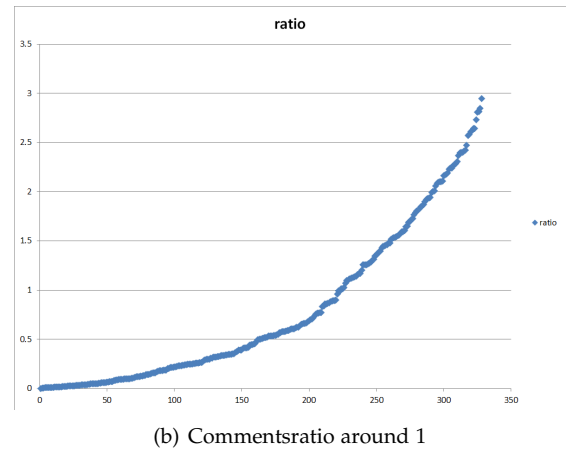
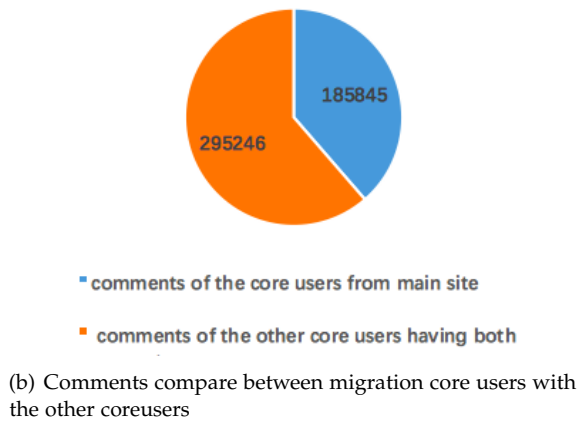
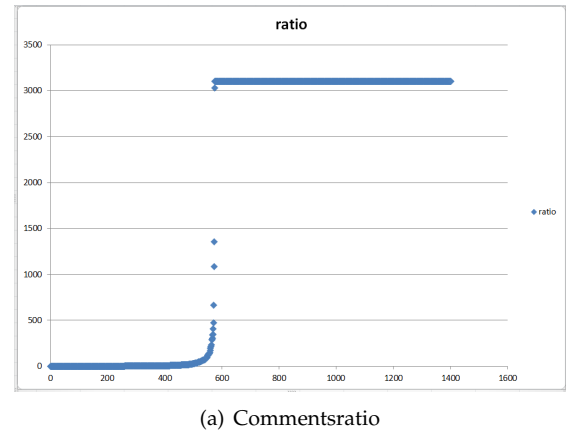
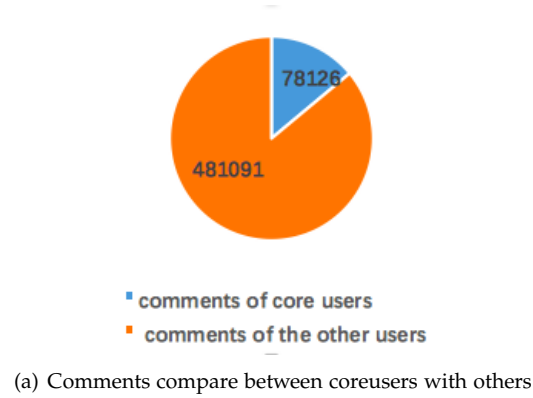
(b)

**Figure 15:** Comment amount of 3247 overlap russian core users in Stackoverflow

For these 3247 overlap core Russian Stackoverflow users , they have 181,643 posts,which includes 59312 questions and 122331 answers. And for those 1402 overlap core Russian Stackoverflow users who from stackoverflow site, they have 77466 posts, which includes 17353 questions and 60113 answers. Compare the average number is , 1:2.062 vs 1:3.464. 1073 of 3247 users never have any posts on Stackoverflow. And 1393 of 3247 users never have any comment in Stackoverflow site. 366 of 3247 users never have both post and comment in Stackoverflow site. Their average scale between post and comment on Stackoverflow is 1:2.175, Their average scale between post and comment on Russian Stackoverflow is 1:1.952.

The comment amount distribution chart is in figure 14. We can see that the majority of the russian users haven't post any comment. And fot that 3247 russian core users who owns two accounts, their comments number is also calculated.

65623 Russain users have 559217 comments(average number is 8.522).4818 coreusers have 481091 comments(average number is 99.853). And the 60805 users who are not core users have 78126 comments(average number is 1.285). 3247 core users in the overlap set have 395088 comments(average number is 121.678). 1402 core users in the overlap from Stackoverflow main site have 185845 comments(average number is132.558) some results in figure 15



**Figure 16:** *comments compare*

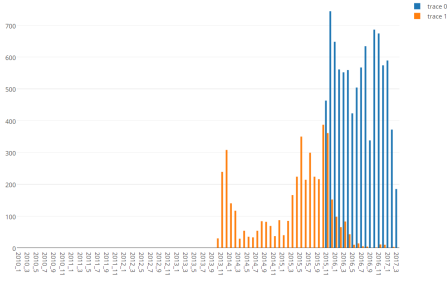
**Figure 17:** *comments ratio*

So the result is the core users from Main site are the most active users in posting comments. And they contributed 33.23% comments. The average comments number of this group of people is 132.558.

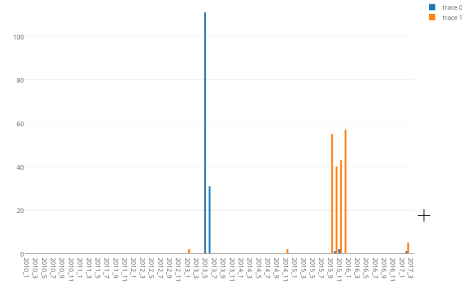
Also, calculating the comment ratio similarly with post ratio. And there are two figure in high similarity.

Similarly, we have those 8 users comments chart that comparing the comment activity before and after they signed up Russian Stackoverflow

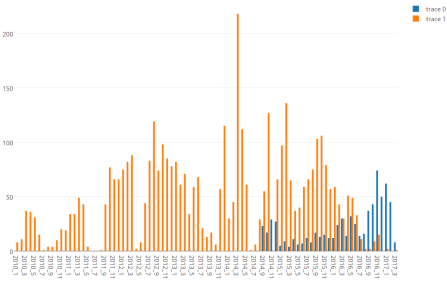




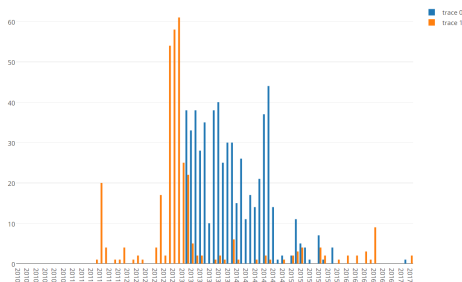
(a) Russian Stackoverflow Active.



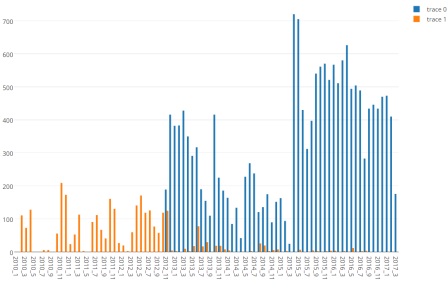
(a) Total Active.



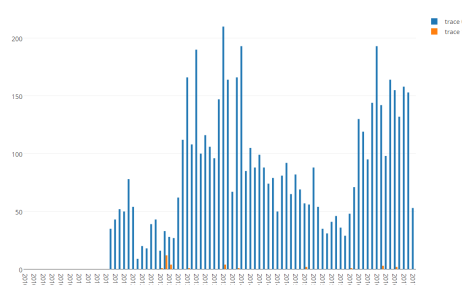
(b) Russian Stackoverflow Inactive.



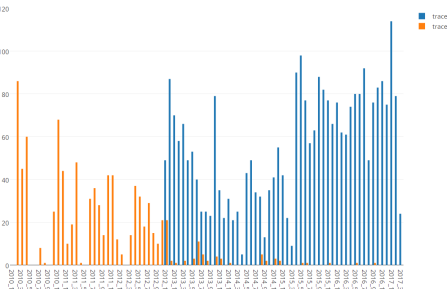
(b) Total Inactive.



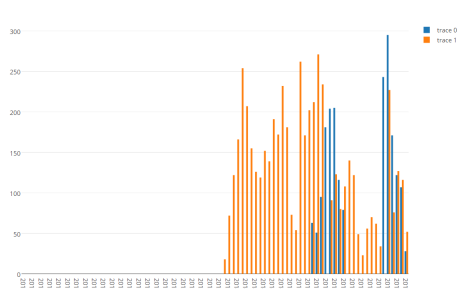
(c) Stackoverflow Active.



(c) Both Active.



(d) Stackoverflow Inactive.



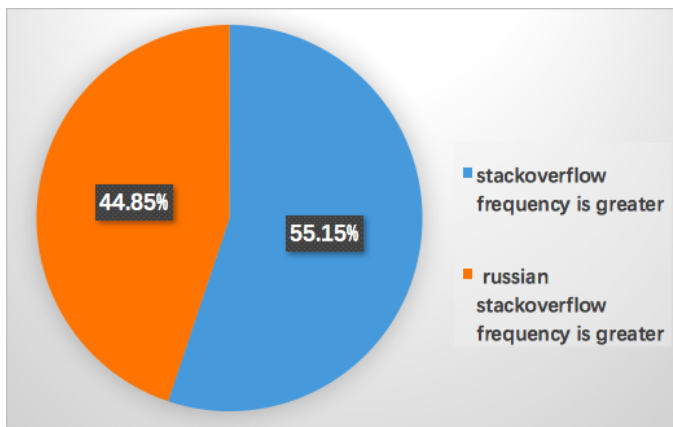
(d) Both Inactive.

**Figure 18: 8 examples(Part 1)**

## VII. TAGS

A tag is a word or phrase that describes the topic of the question. Tags are a means of connecting experts with questions they will be able to answer by sorting questions into

specific, well-defined categories. So in some way, tags of stackoverflow can show some focus areas in which most of users are concentrating. Generally, there are 48373 tags in Stackoverflow site, and 3496 tags in Russian Stackoverflow, which includes 2845 english tags and 651 russian ones. There are 2154 english tags of russian stackoverflow in the overlap set. That is 75.7%. After translating russian to english, those russian tags owns 281 still in stackoverflow tag set, and 114 in russian stackoverflow tags in english, which means there are actually 3382 tags in russian stackoverflow. And we can see the number of unique tags in Russian Stackoverflow is 1061, nearly 31.37%. The total number of using tags is 319186 in Russian Stackoverflow and 39987153 in Stackoverflow. So after calculating the absolute frequency of each tag, we compare the different frequency of the same tag in these two community.



**Figure 19:** *Tags frequency compare*