Is it better to build multi-language Stackoverflow site?

Australian National University
u5870997@anu.edu.com

June 15, 2017

Abstract

I. Introduction

II. Contents

- Link
- User
- Post
- Comment
- Tag

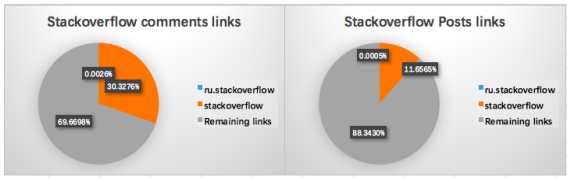


Figure 1: *Links in Stackoverflow*

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 devided into two sets, one is the set of posts and comments who use links to stackoverflowcom and the other is those to ru.stackoverflow.com. Some statistics are presented in figures below.
For the posts and comments in Stackoverflow:

Similarly, for Russian Stackoverflow:

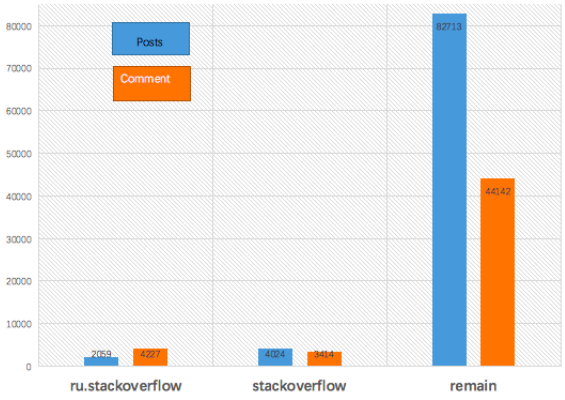


Figure 2: *Amount of links activily in Russian Stackoverflow*

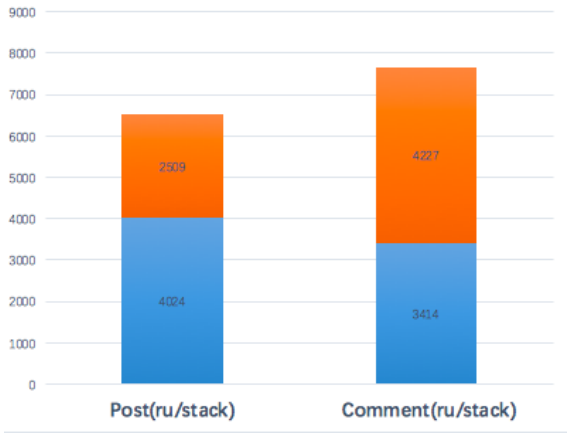


Figure 3: *Compare betweenthe amount of links to stackoverflow site and Russian subsite*

Apparetly, 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 theis 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 Stackoverlfow 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 refenrence 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 structurized 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 th 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 art 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 otehr 7 of these 10 people flow to Russian stackoverflow from Stackoverflow,which means Russian site suprisingly offers a lot of users as feedback to Stackoverflow.

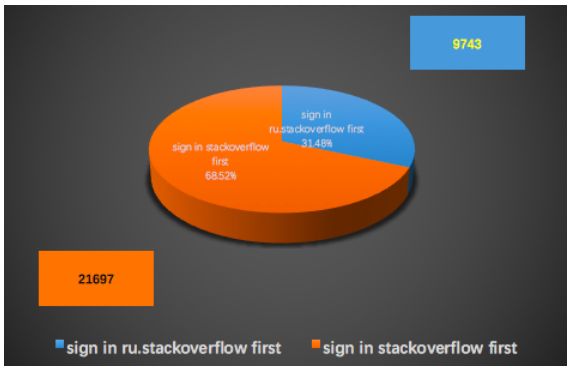


Figure 4: *Creation date compare*

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

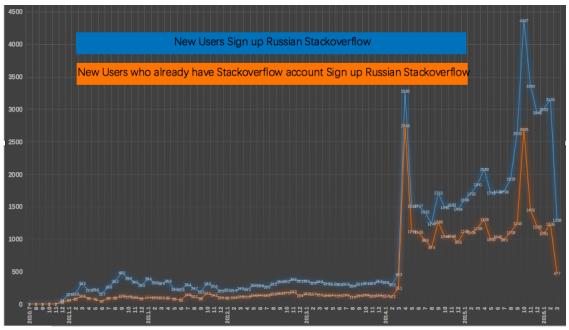


Figure 5: *User movement for signing up Russian Stackoverflow in each month*

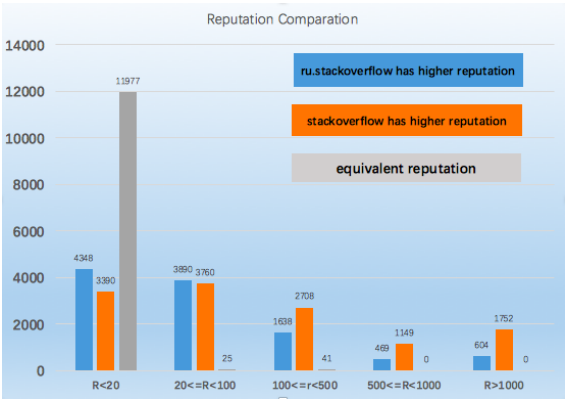


Figure 6: *Reputation compare*

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.

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 foudation 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

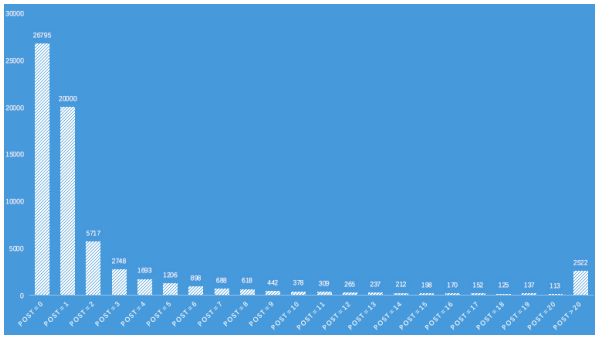


Figure 7: *russian users posts count result*

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 hind,s we need to use their creation date again to defenite their direction of migration between these two sites, which is showed in figure 9.

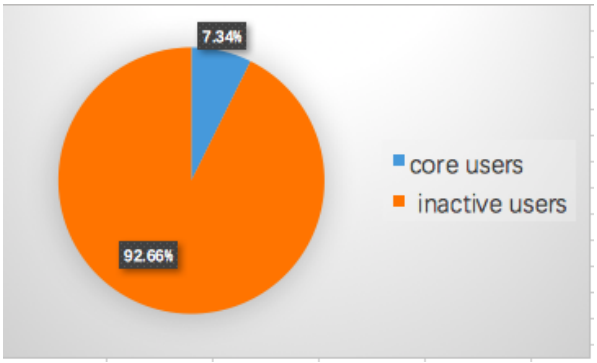


Figure 8: *Core users that owns both Stackoverflow account and Russian Stackoverflow account at the same time*

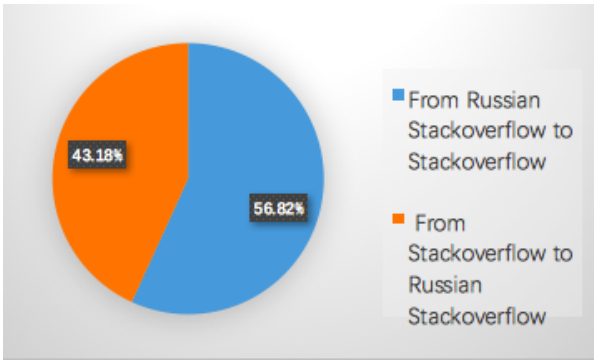


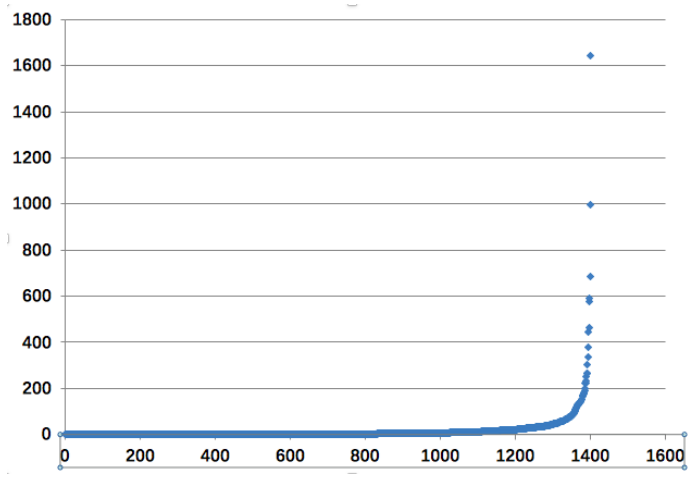
Figure 9: *The compare of migration drection*

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 and 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.

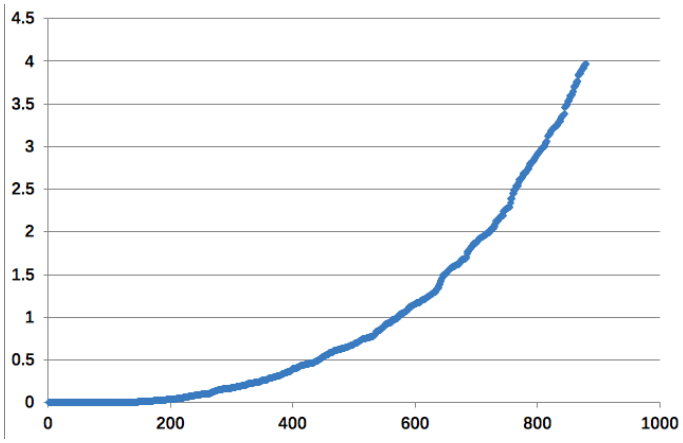
For the 1402 core users who already have Stackoverflow account before thier creation of Russian Stackoverflow account, the compare between the post activities before their migration with the post activities after their migration can definitly reveal their emphasized aspect. The idea is that the object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

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

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

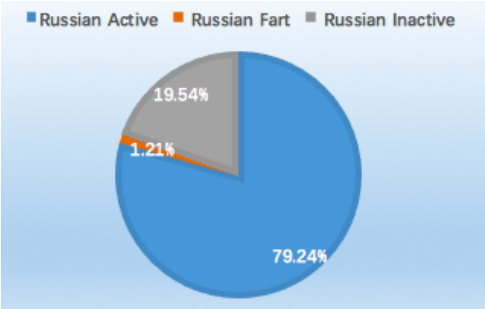


(b) Active ratio close to 1

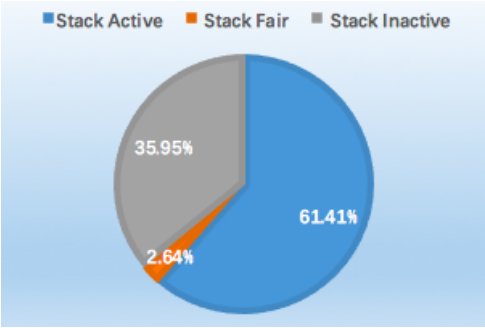
Figure 10: *active ratio chart*

i. PostCount

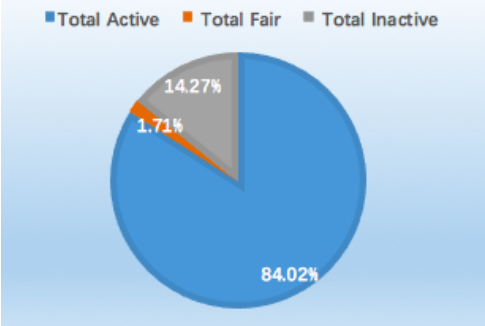
As the result in figure 10, we can easily determine that 0.9 1.1 is the domain of fair activity area. So there are three level of activity, the first one is active, which means the count of users' post is abov 1.1 mutiple of their previous one. The second one is fair, which is between 0.9 and 1.1. The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

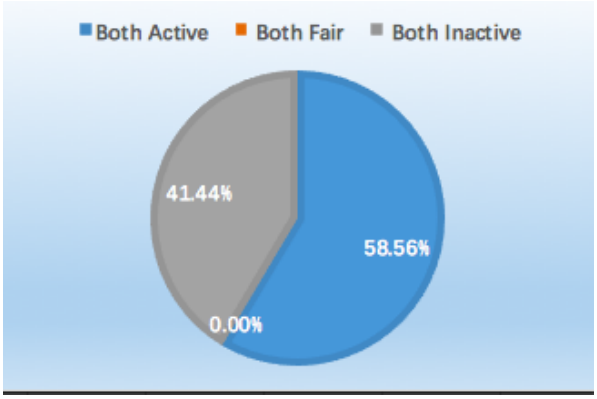
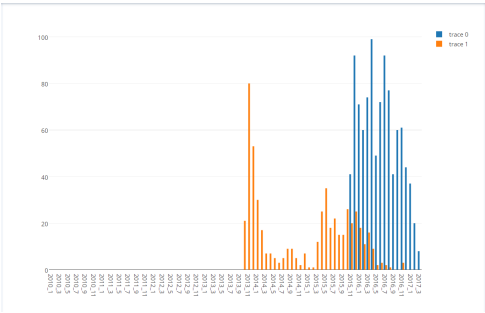
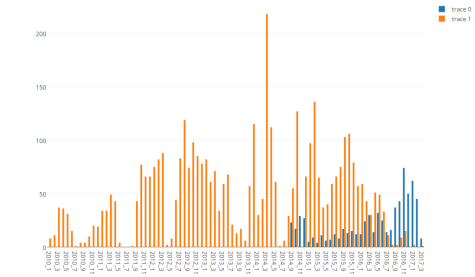


Figure 12: *Overall result*

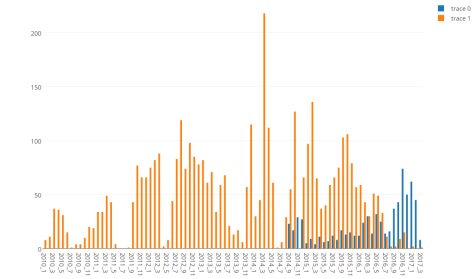
And the in these 1402 users who own both Stackoverflow account and Russian Stackoverflow account and migrated from Stackoverflow site to Russian sub site. 861 of them are both on both sites after their migration, and 3 of them are fair, while 200 of them are inactive. The result is in figure 12.



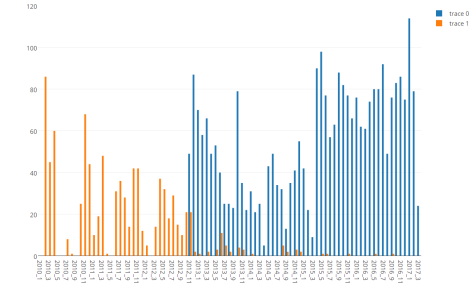
(a) Russain Stackoverflow Active.



(b) Russain Stackoverflow Inactive.

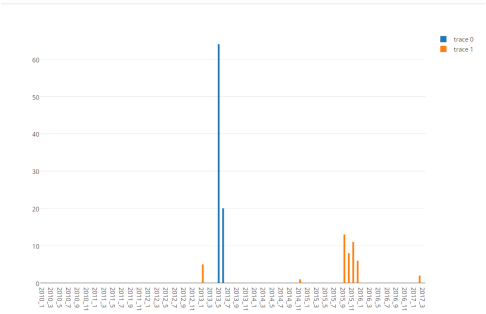


(c) Stackoverflow Activite.

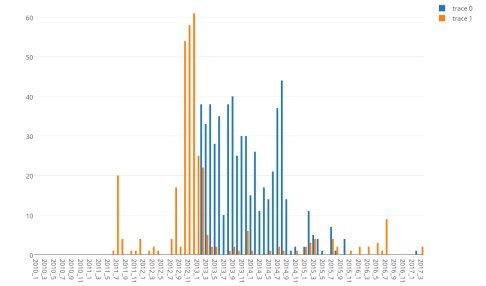


(d) Stackoverflow Inactive.

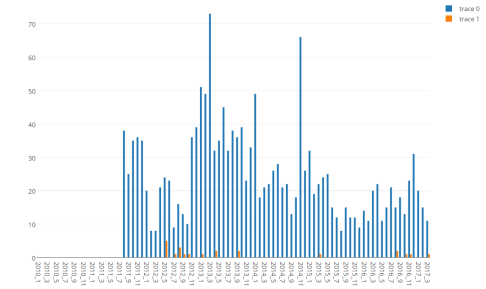
Figure 13: *8 examples(Part 1)*



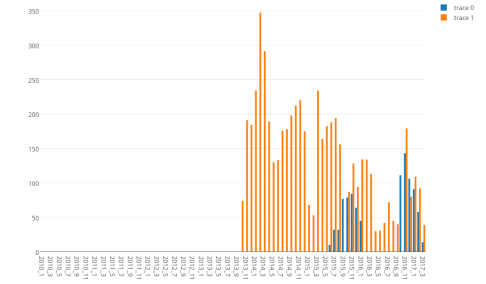
(a) Total Active.



(b) Total Inactive.



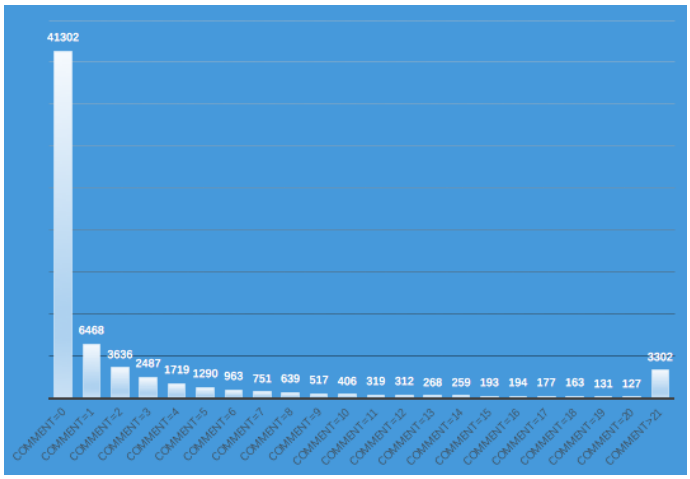
(c) Both Active.



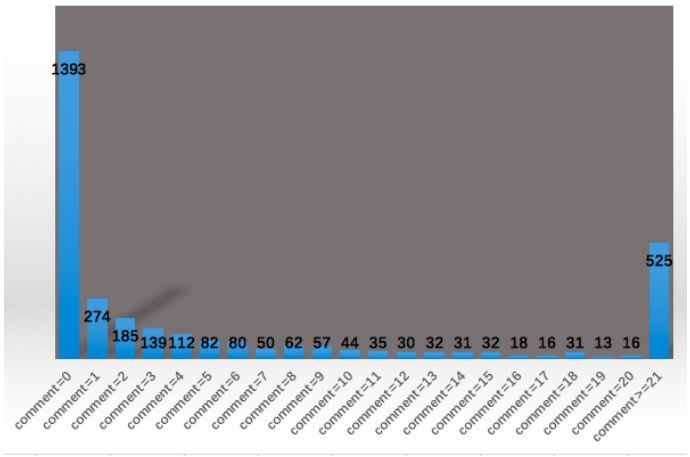
(d) Both Inactive.

Figure 14: *8 examples(Part 2)*

VI. Comments



(a) Comment amount in Russian Stackoverflow



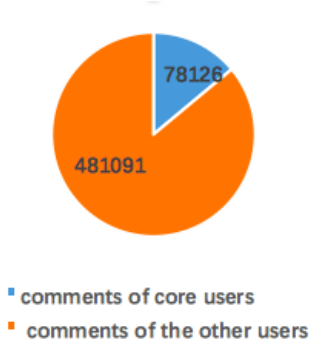
(b)

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

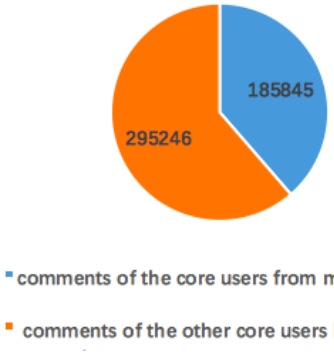
For these 3247 overlap core Russian Stackoverflow users , they have 181,643 posts,which includes 59312 questions and 122331 answers. And for those 1402 overlap core Russian Stackoverflow users who from stackoverflow site, they have 77466 posts, which includes 17353 questions and 60113 answers. Compare 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

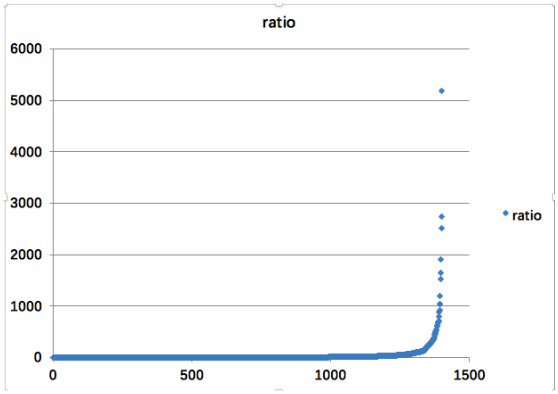


(a) Comments compare between coreusers with others

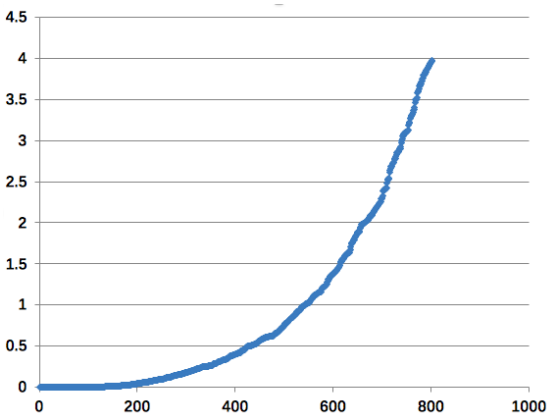


(b) Comments compare between migration core users with the other coreusers

Figure 16: *comments compare*



(a) Commentsratio



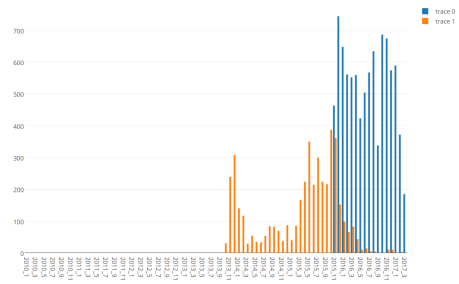
(b) Commentsratio around 1

Figure 17: *comments ratio*

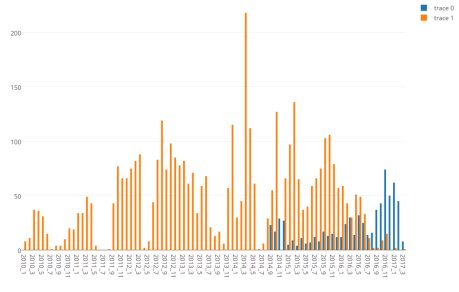
So the result is the core users from Main site are the most active users in posting comments. And they contributed 33.23% comments. The average comments number of this group of people is 132.558.

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

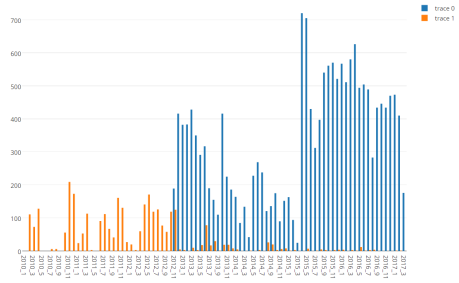
Similarly, we have those 8 users comments chart that comparing the comment activity before and after they signed up Russian Stackoverflow



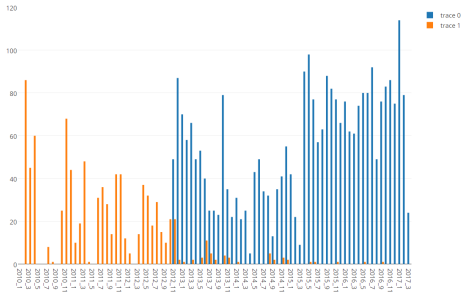
(a) Russain Stackoverflow Active.



(b) Russain Stackoverflow Inactive.

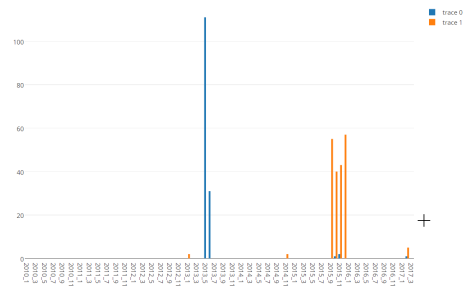


(c) Stackoverflow Activite.

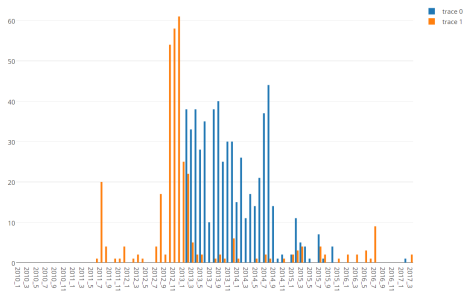


(d) Stackoverflow Inactive.

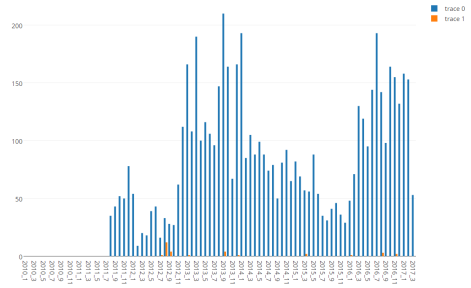
Figure 18: *8 examples(Part 1)*



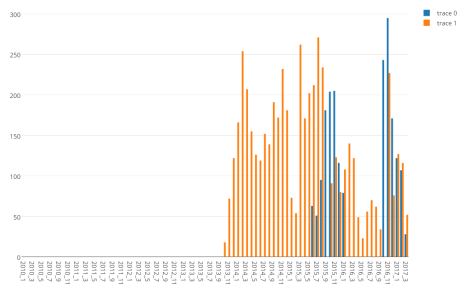
(a) Total Active.



(b) Total Inactive.



(c) Both Active.



(d) Both Inactive.

VII. Tags

A tag is a word or phrase that descibes the topic of the question. Tags are a means of connecting experts with quesitons 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, 114 in russian stackoverflow tags in english, and 87 in overlap set, which means there are actually 3382 tags in russian stackoverlfow. 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.

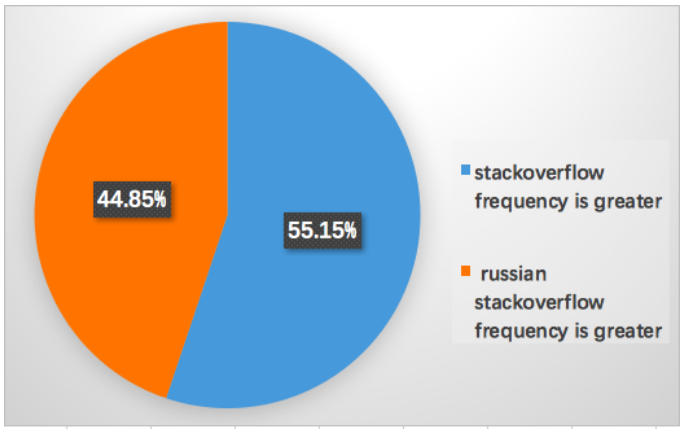


Figure 19: *Tags frequency compare(All english tags)*

Some charts below show the compare of pure frequency calculation.

Id	TagName	Count	ExcerptPo	WikiPostId	frequncy
3	javascript	1339816	3624960	3607052	0.033506
17	java	1223220	3624966	3607018	0.03059
9	c#	1068888	3624962	3607007	0.026731
5	php	1047779	3624936	3607050	0.026203
1386	android	960996	3625001	3607484	0.024033
820	jquery	822436	3625262	3607053	0.020568
16	python	715171	3624965	3607014	0.017885
2	html	632135	3673183	3673182	0.015809
10	c++	502132	3624963	3606997	0.012557
58338	ios	495880	4536664	4536663	0.012401
4	css	455125	3644670	3644669	0.011382
21	mysql	450707	3624969	3607033	0.011271
22	sql	376152	3625226	3607304	0.009407
96	asp.net	311115	3625232	3607037	0.00778
7003	objective-(	274166	3625129	3607044	0.006856
4984	ruby-on-ra	269307	3625098	3607365	0.006735
1	.net	248858	3624959	3607476	0.006223
8	c	244538	3624961	3607013	0.006115
78022	angularjs	223650	9580981	9580980	0.005593
154	iphone	217006	3625240	3607496	0.005427
72	sql-server	193773	3625231	3607030	0.004846
1508	json	189860	4889848	4889847	0.004748
12	ruby	178146	3624964	3607043	0.004455
4452	r	173919	3625322	3607736	0.004349
363	ajax	164813	3877045	3877044	0.004122
46426	node.js	163570	4238969	4238968	0.004091
19	xml	154026	3624968	3607588	0.003852
470	asp.net-m	151584	3625253	3607478	0.003791
58	linux	142033	3625112	3607012	0.003552

(a) The top 30 frequent tags in Stackoverflow.

Id	TagName	Count	ExcerptPo	WikiPostId	frequncy
1176	bootstrap	971	414811	414810	0.003042
1110	vkontakte-	638	416293	416292	0.001999
437	android-sd	569	417176	417175	0.001783
2172	android-fr	457	NULL	NULL	0.001432
47	mvc	443	414815	414814	0.001388
1621	activity	418	617002	617001	0.00131
201	google-ma	348	467741	467740	0.00109
1201	golang	337	413112	413111	0.001056
428	cookie	334	414822	414821	0.001046
335	cms	324	415010	415009	0.001015
534	gui	295	415019	415018	0.000924
244	google	290	416148	416147	0.000909
926	form	284	NULL	NULL	0.00089
410	servlet	246	477012	477011	0.000771
522	network	222	427105	427104	0.000696
644	frameworl	199	422925	422924	0.000623
1499	symfony2	196	438588	438587	0.000614
852	div	194	425844	425843	0.000608
246	yandex-ma	192	433691	433690	0.000602
2410	express.js	167	415065	415064	0.000523
1049	plugin	164	NULL	NULL	0.000514
5574	sublime-te	153	515622	515621	0.000479
9	chrome-ex	148	589742	589741	0.000464
1166	asynctask	148	492029	492028	0.000464
1115	webbrows	142	NULL	NULL	0.000445
625	ui	134	425867	425866	0.00042
405	visual-basi	133	598912	598911	0.000417
2152	qtcreator	127	517514	517513	0.000398
177	bat	121	NULL	NULL	0.000379

(b) The top 30 frequent tags in Russian Stackoverflow.

Id	tagname	Count	ExcerptPostId	WikiPostId	russian_frequncy	translate_tagname
653	список	131	NULL	NULL	0.000410419	list
4120	логирование	129	NULL	NULL	0.000404153	logging
693	браузер	128	464013	464012	0.00040102	Browser
1153	редирект	126	NULL	NULL	0.000394754	Redirect
955	меню	118	NULL	NULL	0.00036969	menu
303	видео	106	NULL	NULL	0.000332094	video
89	установка	97	NULL	NULL	0.000303898	device
86	дерево	95	436451	436450	0.000297632	Tree
701	геометрия	94	NULL	NULL	0.000294499	Geometry
805	чпу	91	469572	469571	0.0002851	CNC
3307	геолокация	90	456614	456613	0.000281967	Geolocation
3476	время	90	NULL	NULL	0.000281967	time
3477	дата	88	NULL	NULL	0.000275701	Date
3756	методы	87	NULL	NULL	0.000272568	methods
357	объекты	86	NULL	NULL	0.000269435	facilities
124	защита	84	NULL	NULL	0.000263169	protection
34	лицензирование	81	520661	520660	0.00025377	Licensing
1428	яндекс	81	NULL	NULL	0.00025377	Yandex
276	роутер	79	427110	427109	0.000247504	router

(c) The top frequent russian tags that exist in both sites

Id	tagname	Count	ExcerptPostId	WikiPostId	russian_frequncy	translate_tagname
77	регулярные-выражения	1897	412987	412986	0.00594324	Regular Expression
192	ооп	940	416054	416053	0.00294499	PIP PLO
141	файлы	858	422721	422720	0.00268809	files
1527	строки	847	609609	609608	0.00265363	Code of the line
265	кодировка	599	425938	425937	0.00187665	Character encoding
242	классы	558	422900	422899	0.0017482	classes
3510	сокет	495	611774	611773	0.00155082	Network socket
459	парсер	459	436457	436456	0.00143803	parser
1789	веб-программирование	459	551717	551716	0.00143803	Web development
6278	ассемблер	455	591825	591824	0.0014255	Assembly language#Assembler
461	вконтакте	443	425840	425839	0.0013879	vkontakte
39	сортировка	400	422891	422890	0.00125319	sort
621	разработка-игр	375	415457	415456	0.00117486	Video game development
154	сеть	360	425942	425941	0.00112787	Network
1092	запрос	356	422898	422897	0.00111534	query
98	книги	334	418037	418036	0.00104641	books
853	функции	314	416047	416046	0.000983752	FUNCTIONS -

(d) The top frequent russian tags only in Russian Site

SSSSS