Patterns of User Participation and Contribution in Global Crowdsourcing: A Data Mining Study of Stack Overflow

Himesha Wijekoon [1] and Vojtěch Merunka [1,2]

[1] *Czech University of Life Sciences Prague, Prague, Czech Republic*
[2] *Czech Technical University in Prague, Prague, Czech Republic*

Abstract

Among many popular crowdsourcing platforms, the Question & Answer website Stack Overflow in Stack Exchange Network is used daily to share knowledge globally by millions of software professionals. Therefore, Stack Overflow data can reveal important patterns in global crowdsourcing beneficial for software industry. The aim of this study was to perform data mining on Stack Overflow data, to discover some of these patterns. Focus of this research was to analyze the global user distribution and contribution. Big data analytic techniques were used for data mining activities using Apache Spark with Python language. Oracle Data Visualization Desktop and scikit-learn python library were used for visualization. The results show that although majority of the users are from USA and India, the average contribution is higher in European countries.

Keywords

Stack Overflow, Data Mining, Big Data Analytics, Crowdsourcing, Software Engineering, User Participation, User Contribution

1. Introduction

Crowdsourcing is basically a type of participative online activity where a person or an organization requests a loosely defined group of people (crowd) to carry out tasks for them using open calls. The crowd undertakes the tasks voluntarily driven by motivation which is not supposed to be financial reasons in all the cases [1]. A new term called Crowdsourced Software Engineering has also emerged to describe the phenomena of using crowdsourcing for various software engineering tasks as it is very popular nowadays [2].

Among many popular crowdsourcing platforms used in software engineering, the Question & Answer (Q&A) website Stack Overflow is used daily to share knowledge globally by millions of software professionals. Therefore, Stack Overflow data can reveal important patterns which will help to get an idea about how software professionals share knowledge in a global scale. Eventually the findings will also help global software companies and crowdsourcing platforms to formulate and re-evaluate their strategies and incentive criteria. The aim of this study is to perform data mining on Stack Overflow data to discover patterns of global user distribution and contribution.

2. Background

Stack Overflow caters wide range of computer programming subjects or topics. In 2015 it has recorded 5.7 billion page views as the number of registered Stack Overflow users was reaching 5 million [3]. The registered users can post questions and answers on the website. All the content is freely

Proceedings of HAICTA 2022, September 22–25, 2022, Athens, Greece
EMAIL: wijekoon@pef.czu.cz (A. 1); merunka@pef.czu.cz (A. 2)
ORCID: 0000-0002-2800-5693 (A. 1); 0000-0002-9056-1439 (A. 2)


available for the public for viewing. It also utilizes a comprehensive reputation management system as Atwood states in one of his blog posts in 2009, that he believes in community moderation [3][4].

Schenk et al. in 2013 in their research has found out that contribution is highest in Europe and North America. Then Asia, which is mostly represented by India; Oceania contributes not as much as Asia, but more than South America and Africa combined. However, they base their research on the transfer of knowledge. Specifically, who (country) raises the question and who (country) answers it [5]. However, it will be beneficial also to perform a comprehensive study on the user distribution across the globe with respect to their contribution and reputation.

Reputation measurement can also be manipulated by users who play around with the gamification methods of Stack Overflow [6]. To tackle this issue, in this research the number of questions and answers posted will be also used to represent the contribution.

When comparing these measurements across users, there is a need of normalization of the figures according to the length of membership for the users. For example, Morrison and Murphy-Hill has used the Reputation per Month without just taking Reputation as the measurement in their research [7]. Similarly, number of answers posted per month and number of questions posted per month can be used in this research in addition to the reputation.

3. Methodology

Methodology of this research is based on the following phases specified by Fayyad et al. for discovering knowledge in databases [8].

3.1. Selection

The public data dump of all user-contributed content on the Stack Exchange Network shared in The Internet Archive is used as the main data source for this research. Following files from Stack Exchange data dump which has been published on 8th December 2017 has been downloaded from The Internet Archive for this study.

- Users.xml (2.36 GB)
- Posts.xml (56.3 GB)

Then the structure of the above xml files were studied to select the most appropriate data items. The Entity Relationship Diagram of the schema is shown in Figure 1.

3.2. Pre-processing

Data mining tasks could not be performed directly on top of downloaded raw XML files due to large file size, flat structure of XML files and unbreakable nature of XML files. Therefore, raw data had to be loaded into another format which Apache Spark can utilize its in-memory processing and parallelization power. A MySQL relational database is used for this purpose. A Python script has been written for each raw XML file which was then executed using spark-submit script which is loaded in Spark's bin directory. The Table 1 shows the number of records loaded into respective MySQL tables.

3.3. Transformation

Conversion of some of the data into appropriate forms was needed before starting data mining activities which are described below.

3.3.1. Extraction of Country Names

Since names of countries/locations have been specified in different formats in raw data, a special Python program was implemented to extract the country name accurately with the help of a free and

open-source Python library named geodict (https://github.com/petewarden/geodict). In the end the location of 1,172,495 users were identified and saved in a new database table. This is 15.83% from all users and 80.24% of all the users who have specified a location.

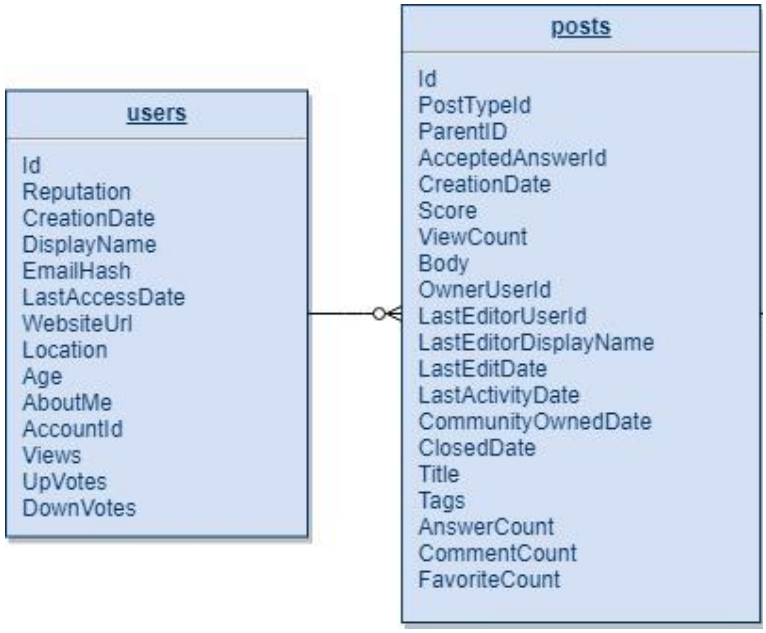

Figure 1: ER Diagram of the Original Schema.

Table 1
Number of Records Loaded into MySQL Tables

MySQL Table Name	Number of Records
Users	7,408,959
Posts	38,360,000

3.3.2. Aggregation

Since tables have millions of data records, Spark with Python API was chosen leveraging the partition aware loading feature. The groupBy function and other built-in aggregate functions like count, avg in Spark were used. All the necessary aggregated data required for the research were generated with the help of Python scripts executed on Spark engine.

3.3.3. Merging

The aggregated data were sometimes needed to be merged prior to data mining. Spark's feature to join RDDs is utilized for this purpose.

3.4. Data Mining

For the numerical data, descriptive summary statistics were used to understand the distribution of data. Mainly the Spark function describe was used for this purpose.

3.5. Interpretation/Evaluation

The descriptive statistics, graphs generated by Oracle Data Visualization Desktop (ODVD) tool and Matplotlib were used to interpret and evaluate the results.

4. Results and Discussion

Country names of 1,172,495 users of Stack Overflow (15.83% from total users) and then 205 country names were identified in the subset under analysis. Top 50 countries sorted in the descending order of user count are presented in Table 2.

Table 2
Top 50 Countries with Users

	Country	Count	Cluster		Country	Count	Cluster
1	UNITED STATES	256470	5	26	VIET NAM	8359	2
2	INDIA	214574	5	27	ROMANIA	8012	2
3	UK	74955	4	28	BELGIUM	7683	2
4	GERMANY	39550	4	29	SWITZERLAND	7406	2
5	CANADA	37576	4	30	ARGENTINA	7277	2
6	FRANCE	30470	4	31	SINGAPORE	7168	2
7	CHINA	30164	4	32	PORTUGAL	7103	2
8	AUSTRALIA	22434	3	33	IRELAND	6906	2
9	RUSSIAN FEDERATION	22070	3	34	DENMARK	6846	2
10	BRAZIL	20070	3	35	SRI LANKA	6508	2
11	PAKISTAN	18661	3	36	JAPAN	6352	2
12	NETHERLANDS	18170	3	37	MEXICO	6327	2
13	INDONESIA	14055	3	38	NEW ZEALAND	6191	2
14	UKRAINE	13391	3	39	MALAYSIA	6179	2
15	POLAND	13027	3	40	TAIWAN	5693	2
16	BANGLADESH	12825	3	41	NORWAY	5475	2
17	SPAIN	12364	3	42	NIGERIA	5288	2
18	PHILIPPINES	12288	3	43	GREECE	5121	2
19	ITALY	12194	3	44	AUSTRIA	5070	2
20	SWEDEN	11928	3	45	COLOMBIA	4765	2
21	IRAN	11862	3	46	SOUTH KOREA	4708	2
22	SOUTH AFRICA	9198	2	47	CZECH REPUBLIC	4405	2
23	ISRAEL	9002	2	48	FINLAND	4251	2
24	TURKEY	8697	2	49	NEPAL	4148	2
25	EGYPT	8527	2	50	BULGARIA	4134	2

As observed United States and India have marginally very high number of users which is more than 200,000 each. Collectively they represent 40% of total users. They are categorized as countries in Cluster 5. Cluster 4 countries have users between 30,000 and 75,000. UK, Germany, Canada, France and China belong to this category. Even though China has the world's highest population, its participation is not matching with the population. It could be due to language issues. This can be same for Russian Federation. Another notable observation is there are only 78 countries with more than 1000 identified users. Cluster 2 represents countries with more than 3000 and only some of them are in top

50 list. Cluster 1 represents countries with less than 3000 users which is not even included in the Table 2.

Above data has been merged with world population data for year 2015 published by United Nations, Population Division [9]. Then users per 1000 capita figure has been calculated for each country for further analysis.

The map in the Figure 2 displays how users per 1000 capita changes across the globe and the Table 3 presents the top 50 countries with users per 1000 capita in descending order. The main observation compared with user count ranking is United States falling to 17th position while India does not even qualify in top 50. However, UK shows consistency in both and the biggest (population wise) country having highest participation. Iceland becomes the number one even though it does not even have sufficient users to be listed in the first list. The main conclusion that can be derived is that most European countries have higher participation per capita generally. The countries like New Zealand, Singapore, Israel, Canada, and Australia are also among the high participating countries.

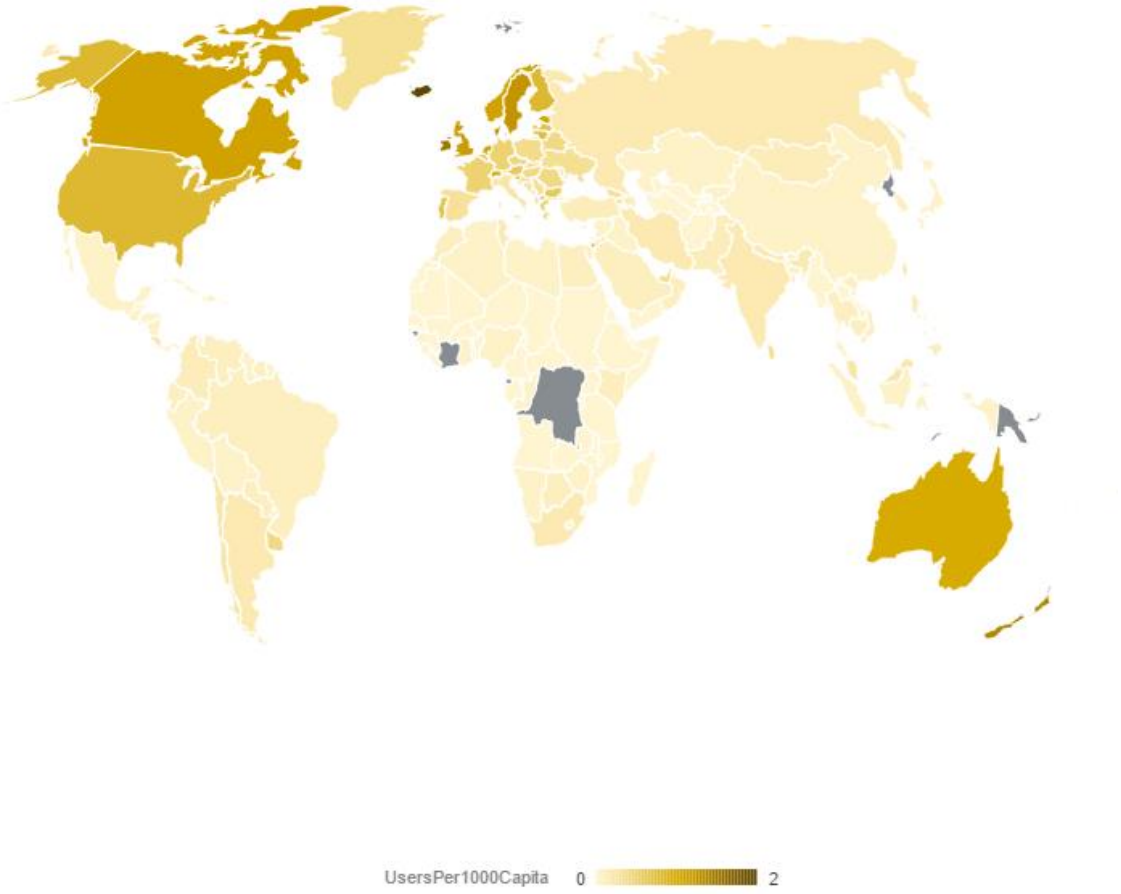

Figure 2: Users per 1000 Capita.

To compare contribution levels of average users of countries, the user contributions in the means of average reputation per user, average number of questions posted per user and average number of answers posted per user from each country have been analyzed. The Table 4 summarizes the rankings of countries which fall into top 20 of each category and has more than 500 users along with Russian Federation and India for their significance. The cells in blue background color displays the ranks within top 20 while cells with pink background displays rankings greater than 20 for the respective category.

As reputation and answer ranking relates to knowledge sharing, respectively Switzerland has become top country in both rankings while closely followed by UK and Germany. Sweden, Austria, and Israel are among top 10 of both rankings with most of other European countries. New Zealand, Austria and Canada contribute much as well.

However, India and Russian Federation have less contribution despite their large population. Another important observation is that most of countries who are reputed, and good answer providers are also good at asking questions. However, Italy, Ireland, Latvia, and Lebanon are basically question askers but not answer providers. Meanwhile Finland, Netherlands and Bulgaria have higher reputation and answering rate, but they do not ask many questions.

Table 3
Top 50 Countries with users per 1000 Capita

	Country	UsersPer1000Capita		Country	UsersPer1000Capita
1	ICELAND	1.91677	26	CROATIA	0.537297
2	MALTA	1.585535	27	CYPRUS	0.484933
3	IRELAND	1.469328	28	GERMANY	0.484042
4	NEW ZEALAND	1.341631	29	FRANCE	0.472717
5	SINGAPORE	1.29497	30	HONG KONG	0.462205
6	SWEDEN	1.221685	31	GREECE	0.456507
7	DENMARK	1.203439	32	MACEDONIA	0.438127
8	UK	1.146152	33	ARMENIA	0.416531
9	ISRAEL	1.116244	34	CZECH REPUBLIC	0.415419
10	NETHERLANDS	1.072704	35	ROMANIA	0.403087
11	NORWAY	1.052918	36	BELARUS	0.395961
12	CANADA	1.045238	37	URUGUAY	0.37942
13	ESTONIA	1.008119	38	HUNGARY	0.372039
14	LUXEMBOURG	0.959874	39	SLOVAKIA	0.359604
15	AUSTRALIA	0.942623	40	POLAND	0.34044
16	SWITZERLAND	0.890169	41	GEORGIA	0.322154
17	UNITED STATES	0.801646	42	SRI LANKA	0.314183
18	FINLAND	0.775452	43	SERBIA	0.312271
19	LITHUANIA	0.718981	44	UNITED ARAB EMIRATES	0.299968
20	PORTUGAL	0.68177	45	UKRAINE	0.299859
21	LATVIA	0.6815	46	COSTA RICA	0.285575
22	BELGIUM	0.680638	47	SPAIN	0.266479
23	SLOVENIA	0.679106	48	BOSNIA AND HERZEGOVINA	0.257921
24	AUSTRIA	0.584192	49	TAIWAN	0.242402
25	BULGARIA	0.575975	50	ALBANIA	0.238767

In both user participation and contribution, European countries along with Israel, Australia, Canada, and New Zealand are highlighted from the rest of the world. These findings were cross evaluated by comparing with the ICT Development Indexes of countries provided by United Nations [10]. The major difference found was the underperformance of crowdsourcing activities of countries like South Korea and Japan who have good global ICT rankings. This situation can be further proven by comparing the findings with the IMD World Digital Competitiveness Ranking 2017 [11]. Even though this must be further analyzed, one reason can be the language barrier. Presence of some other popular alternatives to Stack Overflow also can be also another reason. Under presence of China and Russian Federation can be also due to this.

5. Conclusion

Stack Overflow data reveals important patterns in global crowdsourcing beneficial for software industry. The results on Global User Distribution and Contribution, clearly show that majority of the users are from USA and India. However, in both participation and contribution aspects, European countries along with Australia, Canada and New Zealand have higher rankings. It is also noted the less rankings of Japan, South Korea, Russian Federation, Brazil and China. Since these countries represent huge portion of world population, further studies should be carried out to find factors for this phenomenon.

Table 4
Country Rankings for Contribution

Country	Reputation Rank	Answer Rank	Question Rank
SWITZERLAND	1	1	6
UK	2	4	5
GERMANY	3	3	14
SWEDEN	4	10	13
GUATEMALA	5	55	97
MALTA	6	15	3
ISRAEL	7	2	1
AUSTRIA	8	6	15
NORWAY	9	14	9
NETHERLANDS	10	5	21
AUSTRALIA	11	12	16
NEW ZEALAND	12	13	18
FINLAND	13	11	49
CZECH REPUBLIC	14	7	4
BULGARIA	15	8	38
DENMARK	16	18	7
UNITED STATES	17	22	35
SLOVENIA	18	16	2
CANADA	19	25	24
SLOVAKIA	20	9	20
POLAND	21	17	25
BELGIUM	22	19	10
LATVIA	23	28	17
IRELAND	24	30	11
ITALY	27	23	8
PERU	32	20	55
RUSSIAN FEDERATION	35	38	54
CYPRUS	44	36	19
LEBANON	53	50	12
INDIA	64	58	56

6. References

[1] Y. Zhao, Q. Zhu, 2014, Evaluation on crowdsourcing research: Current status and future direction. Information Systems Frontiers. 2014. Vol. 16, no. 3, p. 417–434.

[2] K. Mao, L. Capra, M. Harman, Y. Jia, 2017. A survey of the use of crowdsourcing in software engineering. Journal of Systems and Software. 2017. Vol. 126, p. 57–84.

[3] Stack Exchange Inc, 2018. About - Stack Exchange. URL: https://stackexchange.com/about.

[4] J. Atwood, A Theory of Moderation - Stack Overflow Blog, 2009. URL: https://stackoverflow.blog/2009/05/18/a-theory-of-moderation/.

[5] D. Schenk, M. Lungu, 2013. Geo-Locating the Knowledge Transfer in Stack Overflow. In: Proceedings of the 2013 International Workshop on Social Software Engineering. Saint Petersburg, Russia: ACM. 2013. p. 2–5.

[6] T. Ahmed, A. Srivastava, 2017. Understanding and evaluating the behavior of technical users. A study of developer interaction at StackOverflow. Human-centric Computing and Information Sciences. 2017. Vol. 7, no. 1, p. 1–19.

[7] P. Morrison, E. Murphy-Hill, 2013. Is Programming Knowledge Related To Age? People.Engr.Ncsu.Edu , 2013. P. 3–6.

[8] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, 1996. From Data Mining to Knowledge Discovery in Databases. AI Magazine. 1996. Vol. 17, p. 37–54

[9] United Nations Department of Social Affairs, Population Division, 2017, World Population Prospects: The 2017 Revision.

[10] United Nations International Telecommunication Union, 2017. ITU | 2017 Global ICT Development Index, 2017. URL: http://www.itu.int/net4/ITU-D/idi/2017/#idi2017rank-tab.

[11] IMD World Competitiveness Centre, 2017, IMD World Digital Competitiveness Ranking 2017, URL: https://www.imd.org/globalassets/wcc/docs/release-2017/world_digital_competitiveness_yearbook_2017.pdf.