International Journal of Engineering & Technology, 7 (3) (2018) 1193-1197

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i3.12489
Research paper

Descriptive and distribution analysis of GitHub repository data

Dinesh Rao [1], Shishir Dubey [1], Mohan Kumar J [1] *, Deepak Rao [1], Balaji B [1]

[1] Manipal Academy of Higher Education, Manipal
**Corresponding author E-mail: mohan.js@manipal.edu*

Abstract

The usage of GitHub by developers is increasing. The organizations are also using GitHub for their project development. As a huge set of users are involved, it leads researchers to analyze the GitHub data. GitHub provides the API (Application programmable interface), to collect its data related to its repository. In our work, the collected data is extensively queried and data is visualized. In this paper, a descriptive analysis is done GitHub data. The results give a lot of insight on the GitHub usage.

Keywords: *Bigdata and Data Analytics; GitHub; GitHub Statistics; Predictive Model; Visualization.*

1. Introduction

The availability of huge data and their applications are growing day by day. The data is available on internet in plenty as open Datasets. Many organizations like World health organisation(WHO), NHS Health and Social Care Information Centre, European Union Open Data Portal, UCLA, National Climatic Data Center and many more[1] provide their surveyed or result data open and online. This data can be used for analysis and visualization. Visualization and analytics on big data enhances the decision making and knowledge discovery [11].

Researchers and scientist prepare their own data by survey or scrapping from internet, when data is not available. Applications like Facebook, Twitter and many others provide API for developers to collect certain set of data and do an analysis on it. In this paper we focus on the descriptive analysis of the GitHub data for various parameters collected through the GitHub API. There are three types of analytics, descriptive, predictive and prescriptive [2]. Descriptive analysis focus on what has happened, predictive analysis is telling the future aspects, what can happen and prescriptive analysis is for how to use data in future. By descriptive analysis, it is possible to find the existing users details and their contributions in their GitHub repository. As various attributes are available, it is important to identify how they affect each other and how they are related with each other. What is the impact of one's attribute to other? As the data is huge, descriptive statistical analysis can be applied.

Found in 2008, GitHub allows us to deposit not only code but also texts and any file which lets us introduce our projects. GitHub also offers the possibility of creating a wiki and a web page for each repository. Also, software for tracking problems. When a developer makes some modifications, those changes are reflected directly to the central repository. According to [5], With Git, if we want to make some variations to a file in a project we can copy the whole repository to our system. We can make modifications on our local copy, then "check in" the changes to the central server. GitHub endorses grittier changes on a project, so it is not necessary to be online or to be connected to the server always whenever any change is happening. Other than hosting services, git adds many of its features, being a command line tool, it also provides the web-based graphical interface.

Copying one user's repository from one user's account to another is a flagship feature of GitHub, which is known as "forking." It allows to get and modify the projects which are inaccessible or don't have write permissions into our account. Users can share the changes being done to the owner using pull requests. These three features including fork, pull request and merge makes GitHub so powerful.

2. Literature review

Descriptive analysis is explored on various applications. Voice recording experiments and descriptive analysis was done with respect to pleasure-displeasure feelings does not correspond to the feelings from natural conversation data [6]. A study on the small and medium-sized enterprises (SMEs) is done [7] and applying descriptive analysis showed that the SMEs in Malaysia are no less different from the rest of their counterparts. One of the interesting insight show the fact, that a good number of SMEs sought Islamic financing modes such as Murabahah, Bai' bithaman Ajil and Ijarah as sources of external capital. This might give positive signal for the Islamic financial institutions to offer more of such facilities to the SMEs. [8] analysis the prescriptive and descriptive analysis for researchers the future direction and strategy. Thousands of data such as patents, reports, web magazines, papers and web news are evaluated. These data are heterogeneous in nature. Descriptive analysis results for the activity history and research power. A group of role model researchers are suggested through the predictive analysis. [9] tries to explore the gap in research that explores automated identification and characterization of expert hackers within online communities. It identifies expert hackers and characterize their specialties by devising a scalable and generalizable framework leveraging two categories of features to analyse hacker forum content. The framework encompasses text analytics for key hacker identification and analysis. An interaction coherence analysis (ICA) framework is used in text analytics, to extract interactions among the users in

Copyright © 2018 Dinesh Rao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

hacker communities as topological feature. Results reveal an interaction network and content based clustering of key actors within the studied hacker community. Descriptive analytics results can be compared between forums to illustrate the commonality and distinctions. [10] identifies and explores the charging profile and preference of electric vehicle(EV) drivers with three main three strands, the charging profile, preferences and recharging facilities. This is done through clustering analyses. The unique nature of the electric mobility (e-mobility) system is investigated with different analysis stages. The study reveals the interest facts on charging patterns and helps to understand the nature of the EV system and plan for new RFs. A predictive analysis and descriptive analysis is done for heart disease diagnosis. A Decision Trees, Naive Bayes, Support Vector Machine and Neural Networks is used for predictive analysis. A descriptive analysis is used for association and decision rules. Better results are shown compared to other methods used [11].

3. Methodology

GitHub API v3 which is a stable version of the API is used in data collection. This API has a limitation on the requests per hour, the limitation is known as the RATE LIMITING. If this API is accessed through basic OAuth, then it allows 5000 requests per hour, either it allows 60 requests per hour if no authentication used. This API uses HTTP redirection where required as well as uses HTTP verbs such as HEAD, GET, POST, PATCH, PUT, DELETE, and this makes it call it as a GitHub REST API. This API is used in the proposed work to collect the user's information, repositories information of the respective users, commit logs of the users on their repositories.

For user's information, 37 fields were extracted using the GitHub API, each for 5000 users. However, only 15 fields are used for quantitative analysis as shown in Table 1. Each user's information extracted is in a CSV format and been appended into a single CSV file which is later stored in MongoDB.

Using GitHub API, 69 fields were extracted from each user's repository, but only 26 fields are being utilized. The fields are shown in Table 2. All repositories information extracted is appended into a single JSON file which is later saved in MongoDB by matching the Id from user's information and owner's id from repository's information to manage the data per user.

Using GitHub API, 8 fields were extracted from each user's repository's commit log, but only 5 fields are being utilized. All repositories, commit logs information extracted, is appended into Multiple JSON files which are later saved in MongoDB by matching the id from user's information and owner's id present in commit field of repository's, commit logs information to manage the data per user.The commit log information is shown in Table 3.

Table 1: GitHub User's Data Fields

Field	Description
Id	GitHub id of a user.
Bio	Information about the user, including Job profile, the field of work and any information what users want to put can put here which can be shared.
Company	User's company.
created_at	The creation date of user's account.
disk_usage	The disk space used by the users for their repositories.
followers_count	The number of followers who are following the user.
following_count	The number of users followed by the user.
last_modified	Last date when the user has made some modifications.
Location	User's new Location and region.
Login	User's login name.
Name	User's new name.
owned_private_repos	Total no. of private repositories owned by a user.
public_repos_count	Total no. of repositories shared publicly.
total_private_gists	Total no. of private gists.
total_private_repos	Total no. of private repositories.

Table 2: User's Repository's Data Fields

Fields	Description
commits_url	URL address for commit logs of the user on a repository.
created_at	Creation date of a repository.
Description	The description of a repository/project.
Language	The language used within the project/repository.
Name	Name of the repository.
default_branch	It shows the branch of the repository.
Fork	The Boolean field which shows whether the respective repository is either forked or not.
forks_count	No. of forks count, which shows the count of repository been forked.
pushed_at	It shows new commit on a repository and updates branch each time the new commit is made.
full_name	Name of the repository with owner's name.
open_issues_count	Count of the issues which are still not resolved.
git_url	URL address of git in a repository.
has_downloads	It shows whether repositories have downloaded or not.
has_issues	It shows whether repositories have some issues or not.
has_pages	It allows to disable or enable pages.
has_wiki	It allows to enable/disable wikis for the repository.
Homepage	GitHub's homepage.
Private	It shows whether the repository is private or not.
Size	The size of the repository in kilobytes.
Id	The id of a repository.
pushed_at	Last date when the files are pushed or commitment made.
stargazers_count	No. of users bookmarking a repository.
updated_at	Last date when modifications being made.
URL	URL address of repository.
watchers_count	Count of users subscribed to "watch" to get notified about project activities.
owners_id	The id of the user to which repository belongs.

Table 3: Commit Logs Fields of A Repository

Fields	Description
author	User to which the repository belongs, on which commits are made.
comments_url	Address of comments on commit log.
commit	Includes all information related to the commits such as author, message, date of commitment
committer	It Includes the name and mail address and information of the committer.
parents	It includes URL for commit logs as well as SHA of the commit.
message	Commit message

4. Results

4.1. Variable identification

Table 4: Identification of Variables

Type of Variable		Data Type		Variable Category	
Predictor Variable	Target Variable	Character	Numeric	Categorical	Continuous
forks_count Size	watchers_count	description language	Id	has_download	forks_count
open_issues_count			forks_count	has_issues	open_issues_count
has_downloads has_pages			Open_issues_count	has_pages	Size
has_wiki			size	has_wiki	watchers_count

Fork has_issues		stargazers_count	Fork language

In this step, Identification of predictors and target takes place, which also includes the identification of data types and category of variables. As shown in Table 4, it contains variables which are been used in the predictive analysis.

4.2. Bivariate analysis

Bivariate Analysis is an analysis of two variables where the correlation between the two variables are taken into an account and are judged according to their association level. Bivariate analysis can be performed on both categorical and continuous variables.
For the bivariate analysis of two continuous variables, scatter plot is considered which is an agile form to find the relationship between two variables. The pattern represents the relationship between variables. The relationship can be linear or non-linear but it doesn't represent the strength of the variables.

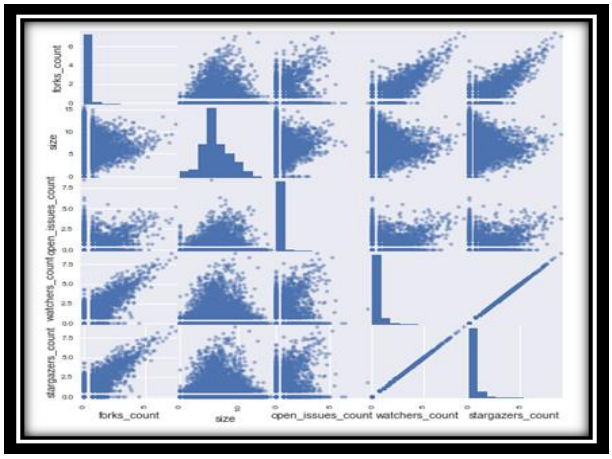


Fig. 1: Relationship between Continuous Variables.

Here in Figure 1 above, the correlation between the variables is shown where relationship holds the correlation such as:

- forks_count holds moderate positive correlation with watchers_count.
- forks_count also hold moderate positive correlation with stargazers_count.
- stargazers_count holds strong positive correlation with watchers_count.

For knowing the strength of these variables, correlation is been used which is derived using the equation (1).

$$\text{Correlation} = \frac{\text{Covariance}(X,Y)}{\sqrt{\text{Var}(X) * \text{Var}(Y)}} \quad (1)$$

Whereas, the correlation varies between the -1 and +1.
-1 shows the perfect negative correlation.
+1 shows the perfect positive correlation.
0 shows the no correlation.
The correlation between the continuous variable is shown below:

	forks_count	size	open_issues_count	watchers_count	stargazers_count
forks_count	1.000000	0.022238	0.459718	0.762623	0.762623
size	0.022238	1.000000	0.033236	-0.014705	-0.014705
open_issues_count	0.459718	0.033236	1.000000	0.429115	0.429115
watchers_count	0.762623	-0.014705	0.429115	1.000000	1.000000
stargazers_count	0.762623	-0.014705	0.429115	1.000000	1.000000

Fig. 2: Correlation of Continuous Variables.

In Figure 2., above, the good positive relationship is there between the variables which are 0.76 between stargazers_count and watchers_count and even others too are less correlative but no variables hold zero correlation, which is actually good for making further analysis.
For the Bivariate analysis of two categorical variables, there exist so many techniques such as Two Way Table, Box Plot Visualization and etc, Barplot visualization is been used in this project to find the relationship between categorical variables which is shown in Figure 3.



Fig. 3: Relationship between Categorical Variables.

Here we can see above in Figure 3, the variables which are having good relationships are fork, has_issues and has_wiki which holds a good relationship and can be used for further analysis.

4.3. Descriptive analysis and visualization

In the proposed work a descriptive Analysis is carried out. As the term implies, it is used to describe the data or summarize the data in a human-interpretable form which uses data aggregation and data mining to grasp the insights of past and present the fact – "what has happened?", the past can be referred to any point of time that event has occurred, whether that has occurred a while ago or some years ago. The descriptive analysis allows learning from the past and how it might affect future outcomes.
Most of the statistics such as (Max, Min, Sum, Averages, Percent changes) comes under this category. There are an infinite number of these statistics available and are widely used for financial metrics which is commonly reported, is a product of descriptive analysis. For example, Year over year pricing changes, month over month sales growth, the number of customers and total revenue per subscriber, all these features describe what has happened in the business in the total period being measured.
The descriptive statistics on 5000 user's information and 25885 repositories information is being shown in Figure 4 and Figure 5.

	followers_count	following_count	public_repos_count
count	4998.000000	4998.00000	4998.000000
mean	6.448780	4.02561	7.303521
std	47.209449	37.09562	22.259403
min	0.000000	0.00000	0.000000
25%	0.000000	0.00000	0.000000
50%	0.000000	0.00000	1.000000
75%	2.000000	1.00000	6.000000
max	1744.000000	1848.00000	776.000000

Fig. 4: Descriptive Statistics of 5000 User's Information.

Figure 4 shows the descriptive statistics on the selected fields of Users Information which show that followers count varies from 1 to 1744 and on the other hand, the following count varies from 1 to 1848 with the public repositories varying from 1 to 776. Descriptive statistics on repositories information is being shown in Figure 5.

	forks_count	size	owners_id	open_issues_count	stargazers_count	watchers_count
count	25885.000000	2.588500e+04	2.588500e+04	25885.000000	25885.000000	25885.000000
mean	1.032838	1.332051e+04	4.325071e+05	0.609735	4.007031	4.007031
std	17.979457	9.874025e+04	3.275880e+05	31.151710	76.553956	76.553956
min	0.000000	0.000000e+00	1.423000e+03	0.000000	0.000000	0.000000
25%	0.000000	1.120000e+02	1.613590e+05	0.000000	0.000000	0.000000
50%	0.000000	2.800000e+02	3.911900e+05	0.000000	0.000000	0.000000
75%	0.000000	2.360000e+03	6.845590e+05	0.000000	1.000000	1.000000
max	1687.000000	6.591493e+06	2.245887e+07	4916.000000	7793.000000	7793.000000

Fig. 5: Descriptive Statistics of 25885 Repositories.

As here in Figure 5 above, it shows the descriptive statistics on the selected fields of Dataset. Here, we can see that the forks_count or the downloads of the project varies from 0 to 1688 and the size of the project ranges from 0Bytes to 6.59 Mbytes and the open issues count ranges from 0 to 4916. Here, the range of stargazers as well as watchers is similar which ranges from 0 to 7793. As here the stargazers count and watchers counts hold the same results which signify the strong relationship between those variables.

As to know more about the insights of GitHub data, the data exploration got started by seeing the distributions of some fields like public_repos_count, followers_count, followees count, forks_count, open_issues_count,

Note: As Data was collected randomly, so there are so many new users too which don't have an even single repository yet but they are also considered as a part of the project. So, all the results gained in this project are true according to the collected data and the plots showing the distributions are all logarithmic normalized with log(x+1) to overcome the high variance in the values.

4.4. Distribution analysis

Distributions are used to reduce the computational cost of the method such as regression or classification which means, later at some point, distribution is to be estimated from any data which is helpful while making Predictive Analysis and gives a better idea about the data. So here are some distributions of GitHub (DG) data are presented, which gave some good idea about the data.

DG1: Distribution of Public Repositories

The distribution of the public repositories is being shown below in Figure 6, which is displaying the percentage of users and their public repositories. According to the collected data, 40 % of users have no repository yet .and almost 32 % users have at least one public repository and the remaining users have public repositories varying from 2 to 62.

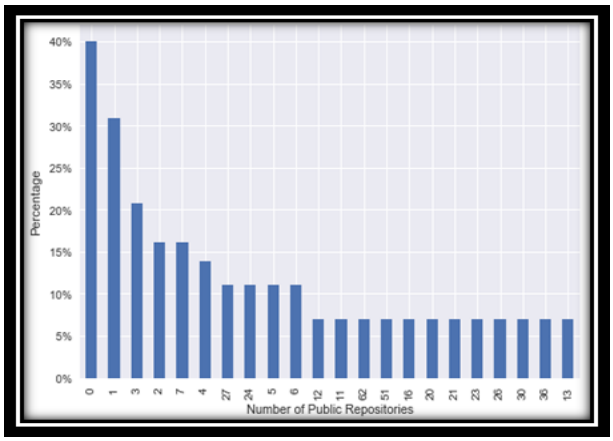


Fig. 6: Distribution of Public Repositories.

DG2: Distribution of Followees

This distribution of followees shows the percentages of followees present within the data. The distribution for Followees in Figure 7 Fig is showing that the less percentage of users are following many users, but the large percentage of users are following few users or even not following any of the users. As we can see that, 40 % of users are not following any of the users and 32 % of users are following at least two users and other percentages of users are following other users which vary from 1 to 99.

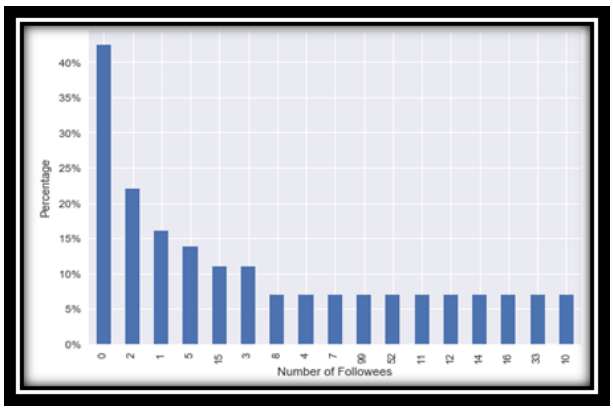


Fig. 7: Distribution of Followees.

DG3: Distribution of Followers

This distribution of followers is showing the percentages of users having followers. As here in Figure 8, we can see that the less percentage of users have a large number of followers but the large percentage of users have less number of followers or even one follower. It is clearly visible in the figure that more than 40 % of users have no followers, and more than 30 % of users have at least one follower and the remaining percentage of users have number of followers which varies from 2 to 684 followers.

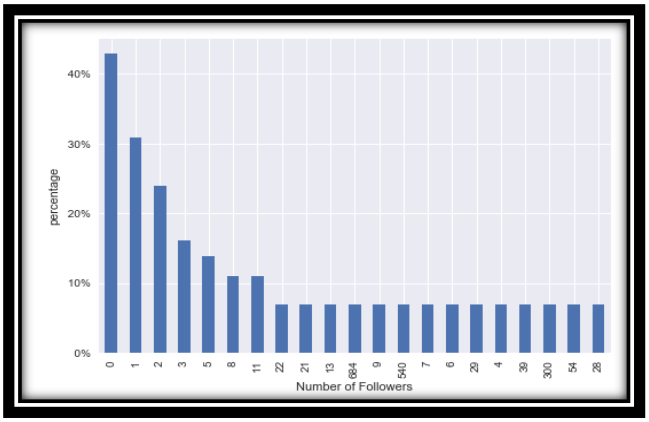


Fig. 8: Distribution of Followers.

DG4: Distribution of Forks
As in Figure 9, it shows the distribution of forks for every project which shows the number of times the projects were cloned or forked. It is shown both in statistics table and figures that most of the projects have fork count equal to zero. We can easily see that the large no. of projects are not forked but the projects which are forked many times are very less in number.

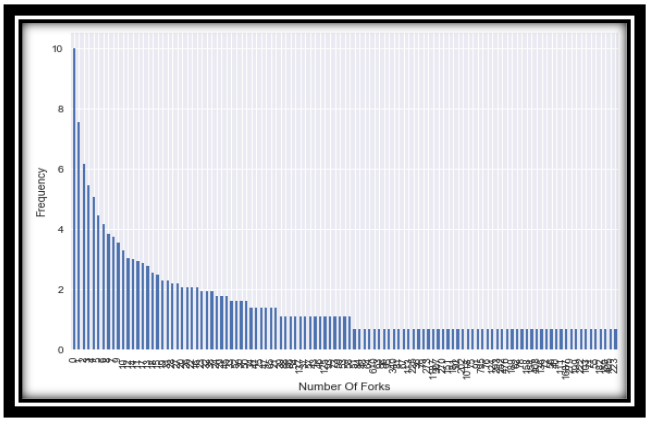


Fig. 9: Distribution of Forks.

DG5: Distribution of Open Issues
Here the distribution of open issues count has been shown below in Figure 10, which shows the frequency of projects with the open issues. As the open issues are the projects which are open and having issues, project issues are open to both owners as well as to the public. As here the distribution shows that most of the projects have either no issues or 1 issue and other issues vary from 2 to 4916. Here, most of the projects are not having issues and projects which are having issues are very less in numbers.

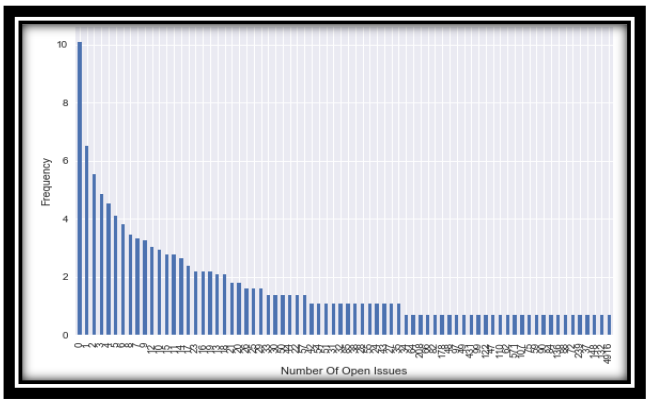


Fig. 10: Distribution of Open Issues.

DG6: Distribution of size of repositories
The distribution of the size of the repositories shows the variance in sizes of repositories. As we can see in Figure 11, more than 60 % of projects are either empty or of 48 kilobytes in size and the remaining percentages of projects sizes are varying from 3 to 240 kilobytes in size.

5. Conclusion

The GitHub data is collected through the API provided and mined using MongoDB. Descriptive analysis and its visualization is done. A descriptive and distribution analysis is done on the collected fields. Visualization for the same is carried out. Many insights are drawn by considering different variable field of GitHub repositories.

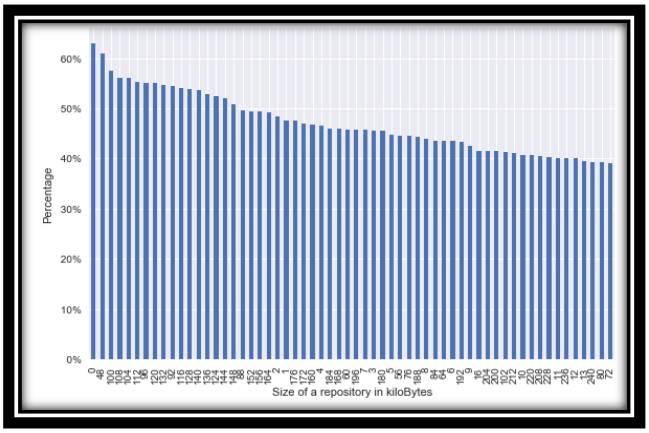


Fig. 11: Distribution of Size of Repositories.

Acknowledgement

We acknowledge Dr. Deviprasad for his constant inputs and suggestions in completion of the work.

References

[1] Forbes, Bernard Marr, (December 2016), https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/#19b807dfb54d
[2] Hadoop Illuminated, (2018), http://hadoopilluminated.com/hadoop_illuminated/Public_Bigdata_Sets.html
[3] Denzyre, (08 Feb 2016, Last Update Made On January 22, 2018) https://www.dezyre.com/article/types-of-analytics-descriptive-predictive-prescriptive-analytics/209
[4] Halo Business Intelligence, (2018), https://halobi.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/
[5] K. Finley. (2012). "What Exactly Is GitHub Anyway?," TechCrunch, 14-Jul-2012. [Online]. Available:https://techcrunch.com/2012/07/14/what-exactly-is-github-anyway/. [Accessed: 20-Apr-2017].
[6] Shimura, Makoto, Fumiaki Monma, Shunji Mitsuyoshi, Masaki Shuzo, Taishi Yamamoto, and Ichiro Yamada.(2010), "Descriptive analysis of emotion and feeling in voice." In Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on, pp. 1-4. IEEE. https://doi.org/10.1109/NLPKE.2010.5587794.
[7] Ab Manan, Siti Khadijah, Jaizah Othman, and Asmady Shahadan. (2011), "Descriptive analysis on the pattern of SME financing in Malaysia." In Sustainable Energy & Environment (ISESEE), 2011 3rd International Symposium & Exhibition in, pp. 139-147. IEEE. https://doi.org/10.1109/ISESEE.2011.5977122.
[8] Song, Sa-Kwang, Donald J. Kim, Myunggwon Hwang, Jangwon Kim, Do-Heon Jeong, Seungwoo Lee, Hanmin Jung, and Wonkyung Sung. (2013), "Prescriptive analytics system for improving research power." In Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on, pp. 1144-1145. IEEE. https://doi.org/10.1109/CSE.2013.169.
[9] Abbasi, Ahmed, Weifeng Li, Victor Benjamin, Shiyu Hu, and Hsinchun Chen. (2014) "Descriptive analytics: Examining expert hackers in web forums." In Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint, pp. 56-63. IEEE.
[10] Elbanhawy, Eiman Y. (2014) "Descriptive analysis of e-mobility system." In Environmental Friendly Energies and Applications (EFEA), 2014 3rd International Symposium on, pp. 1-7. IEEE. https://doi.org/10.1109/EFEA.2014.7059941.
[11] Hussain, Shujaat, and Sungyoung Lee. (2015), "Visualization and descriptive analytics of wellness data through Big Data." In Digital Information Management (ICDIM), 2015 Tenth International Conference on, pp. 69-71. IEEE. https://doi.org/10.1109/ICDIM.2015.7381878.
[12] Babič, František, Jaroslav Olejár, Zuzana Vantová, and Ján Paralič. (2017), "Predictive and descriptive analysis for heart disease diagnosis." In Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on, pp. 155-163. IEEE. https://doi.org/10.15439/2017F219.