How GitHub Contributing.md Contributes to Contributors

Naoki Kobayakawa
University of Tsukuba
Tokyo, Japan
naoki.kobayakawa@gmail.com

Kenichi Yoshida
University of Tsukuba
Tokyo, Japan
yoshida.kenichi.ka@u.tsukuba.ac.jp

***Abstract*—to make open source software development successful, acquiring contributors from outside is an important factor. However, despite the need for external contributions, many projects have disappeared without being able to acquire contributors. GitHub, an Internet hosting service for the purpose of open software development, is widely used for research projects of related topics. However, different from the original research purpose, many pseudo-projects use GitHub for other purpose such as a free file storage. There existence disturbs the analysis of research projects. It is necessary to exclude them. In this research, we focused on the contribution file (contributing.md / contributing.txt) which is the common file of GitHub project. We use contribution file to judge whether the project intend social coding depending on its existence. We also plan to analyze the effect of its contents on acquiring contributors. To do so, we extract information of 459 projects from GitHub archive, and analyze the contents of contribution file and relationship between the contents and acquired contributors.**

Keywords GitHub; contributor; open source software; software development

I. Introduction

Open Source Software (OSS) has been adopted not only for the personal purpose software products but also for core systems of companies and public institutions. It becomes indispensable for our society. In order to evolve and grow OSSs, it is important to acquire not only stakeholders but also contributors widely from outside. Especially, in the case of large OSS, a large number of contributors are needed. Therefore, it is extremely important to explore the acquisition mechanism of contributors.

In this research, we try to elucidate the mechanism by analyzing the OSS projects on GitHub. Although GitHub is an Internet hosting service aimed at supporting software collaborative development, there are many other uses that actually do not require contributors, such as free file storage or learning of version control system. To analyze collaborative software development projects, it is necessary to exclude projects other than the purpose.

To exclude unrelated use projects, we pay attention on so called contribution file. GitHub has a common file called contribution file (contributing.md/contributing.txt [1]) to describe the necessary contribution for the project. The file describes the contribution method that the project expects so that the applicant can judge whether the applicant participates by seeing the contents. By extracting the project including this file, we are able to analyze the projects which intend collaborative software development. In this research, 459 projects including contribution file were extracted from a GitHub archive. By analyzing these projects, we try to clarify whether contributors actually decide to participate by checking the file contents. Also we try to analyze what kind of description is important to acquire contributors.

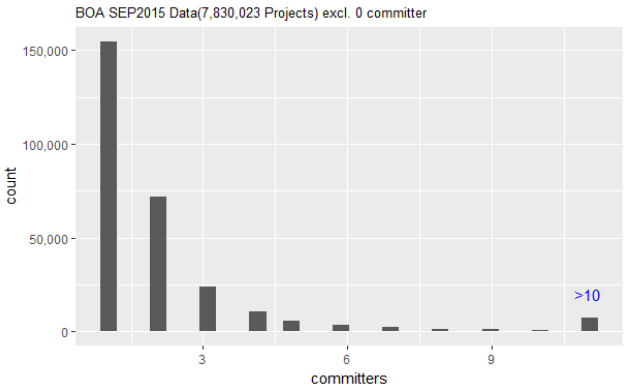

Fig. 1. Number of Committer

II. Related Work

As of April 2017, GitHub has a 57 million projects [2]. But there are very few projects that include contribution file. Also very few projects has enough contributors (See Fig.1).

Tsay et al[3] show that projects with “contributing.md” are 25% to 45% more likely to be alive (7.2% vs. 5.0% for projects that had commit within 7 days, 14.8% vs. 11.8% for 30 days). However, it is a study on the existence of contribution file. It does not extend its analysis to the contents. Chen et all [4] report that clear contribution description is a significant factor to attract new contributors. However, they do not describe the specific characteristics of description which attract contributors.

Based on these related works, we try to analyze relationship between the contents of contribution file and acquired contributors.

0730-3157/17 $31.00 © 2017 IEEE
DOI 10.1109/COMPSAC.2017.139

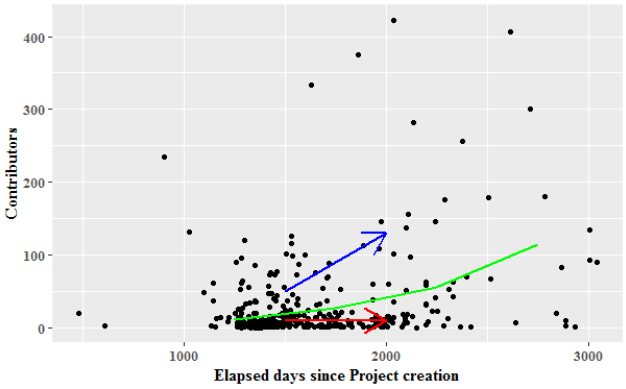

Fig. 2. Growth of Contributors

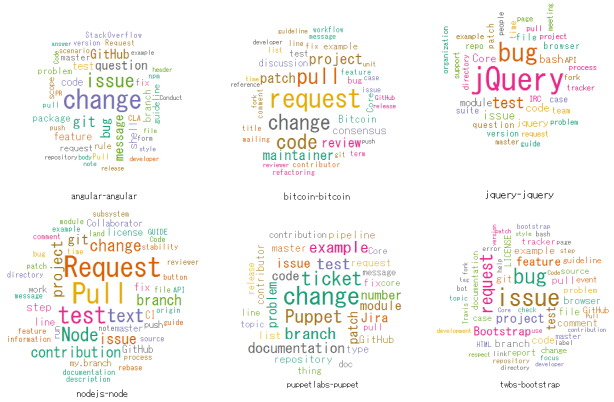

Fig. 3. Word Cloud of Major GitHub Project

III. Data Preparation

In this research, we use data of BOA [5] provided by Iowa State University which is a GitHub archive (September 2015 Full) to efficiently discover projects that contain contribution files. The proportion of projects that contain contribution files is extremely low, and if we search directly from the web, it takes a lot of time. BOA is a huge archive of GitHub Project which adopted Hadoop's technology, and it can acquire the target Project and its attributes in a very short time. We searched 7,830,023 projects and found that as of September 2015 only 639 Projects contain contribution file. Since there is no contribution file on the BOA, we obtained them by using the acquired URL, but as of April 2017 there were only 459 files. After all, in 7,830,023 projects, we were able to get only 459 contribution file. (0.006%)

Project attributes are obtained from projects where contribution file exist by using GitHub API. Attributes acquired are mainly items that can be checked on the GitHub web screen such as Stargazer number (Bookmark), Subscriber number and so on. In order to track changes over time, we plan to continue to acquire attributes every month.

IV. Study Plan

Example of the data obtained from GitHub is shown in Fig.2. The X axis is the number of days elapsed since project creation. The Y axis is the number of contributors. The green line is the average contributors. In the vicinity of 1000 days, the number of contributors is small, but it becomes a bipolarization pattern. Blue arrow gradually acquires contributor, and grows over time. Red arrow disappears with little contributor. We are interested in the mechanism that makes this difference.

The contribution file contents differ according to the projects. Fig.3 shows Word Cloud of the contribution file of the top ranking GitHub projects [6]. As the frequency of occurrence increases, the font size of the word becomes larger and closer to the center. We plan to analyze the relation between this contribution file contents and the number of contributors mentioned above.

In this research, we plan to analyze followings.

1) The correlation between number of contributors and the contents of contribution files.
2) Relevance of maturity of projects and contribution file contents
3) Model to predict the number of contributors in the future.

We plan to use following method to analyze data.

1) Extract words from Contribution.md and create a co-occurrence matrix (bag-of-word) between documents and words. Execute deletion of stop word, stemming, TF-IDF as preprocessing. We classify the results by using non-hierarchical clustering method such as k-means and compare the difference in the number of contributors of each cluster.
2) Next, using a topic model method such as LDA, it breaks down into documents - topics - words. Then, using topics - documents matrix, clustering and regression analysis are carried out and the relationship with the number of contributors is examined. Topics are assumed to be Question, Bug-report, New-feature, Change-rules, Workflow, Coding-guide, License policy.
3) The above two methods analyzes based on the number of occurrences of words, and does not take into consideration the connection and appearance order of words. Using a method such as N-gram or Doc2Vec, vectorize the entire document and analyze it.

V. Future Research

The attributes acquired by the API is information that can be directly referenced from the Web, such as Stargazer number (Bookmark), Subscriber number (Mailing List), and so on. The success criteria of project is not just the number of contributors. Indicators of activity level such as number of commits and pull requests are also important [7]. We would like to include them in the analysis.

Contributors also contribute different purposes and profiles [8]. We would like to investigate that what kind of contributors tend to be acquired based on the contents of contribution files.

REFERENCES

[1] Https://github.com/blog/1184-contributing-guidelines.

[2] "Celebrating nine years of github with an anniversary sale," https://github.com/blog/2345-celebrating-nine-years-of-github-with-an-anniversary-sale, Github. Retrieved 2017-04-11.

[3] J. Tsay, L. Dabbish, and J. Herbsleb, "Influence of social and technical factors for evaluating contribution in github," in *Proceedings of the 36th international conference on Software engineering*. ACM, 2014, pp. 356–366.

[4] J. Chen and I. Portugal, "Analyzing factors impacting open-source project aliveness."

[5] R. Dyer, H. A. Nguyen, H. Rajan, and T. N. Nguyen, "Boa: A language and infrastructure for analyzing ultra-large-scale software repositories," in *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, 2013, pp. 422–431.

[6] Https://github-ranking.com/.

[7] C. Izquierdo, V. Cosentino, and J. Cabot, "Attracting contributions to your github project," *The Journal of Object Technology*, 2015.

[8] L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb, "Social coding in github: transparency and collaboration in an open software repository," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012, pp. 1277–1286.