

StackOverflow and GitHub: Associations Between Software Development and Crowdsourced Knowledge

David Arroyo Menéndez

March 8, 2020

Title: StackOverflow and GitHub: Associations Between Software Development and Crowdsourced Knowledge

- Aleksander Serebrenik
- Bogdan Vasilescu
- Vladimir Filkov

URL:

<https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6693332>

Citations: 203 Published on: 2013 International Conference on Social Computing (IEEE)

Abstract

Big Problem

Associations Between Software Development and Crowdsourced Knowledge

Small Problem

Stackoverflow role in the work cycle of open-source developers

Common Approach

StackOverflow activities and the development process, reflected by code changes committed to the largest social coding repository, GitHub.

One Problem

Approach

Two Conclusions

Introduction

Context of the Work

- Description about StackOverflow and Github

Problem that is addressed

Goal: Is participation in StackOverflow related to productivity of GitHub developers?

RQ1: (Macro level) How are the overall activity levels of developers related across the two platforms? E.g., do active GitHub committers ask more or fewer questions on SO? Are more active answerers also committing more?

RQ2: (Intermediate level) Are the commit (work) patterns on GitHub of developers more active on SO different than the commit patterns of those developers less active on StackOverflow?

RQ3: (Micro level) Is there a functional interaction, or coordination between commit and question/answer activities? I.e., when commits are close to Q&A in time, are they more frequent? How about vice-versa?

- participating in StackOverflow leads to interruptions that could impair a developer's performance
- participating in StackOverflow speeds up development activities
- “good” questions and to give “good”

answers

Data preparation: Stackoverflow and Github

Extraction

- Stack Exchange data dump
- GHTorrent

Preprocessing

- In this paper we consider only the commits

which record the same person as both author and committer (97.8% of the commits in our dataset), and record the date at which a commit was authored (rather than committed).

- Restrict our study to the period July 2011 to April 2012
- We ignore duplicate commits, i.e., commits authored by the same person and having the same timestamp

Intersecting the two datasets

- A key step in our process is merging the GitHub and StackOverflow

Macroscopic View

To study distributions of the number of events of each type (C for commit, Q for question, A for answer).

Comparing Multiple Distributions

The first step is commonly carried out by means of ANOVA or its non-parametric counterpart, the Kruskal-Wallis one-way analysis of variance by ranks. The second step uses the t-test or the rank-based Wilcoxon-MannWhitney test, with Bonferroni correction

Results

- Active GitHub committers ask fewer questions on StackOverflow than others.
- More active GitHub committers provide more answers on StackOverflow
- The least active askers author more commits than others.

Methodology

- To evaluate the committing rhythm of a developer, we calculate the Gini index over the lengths of her inter-commit time intervals.

Results

- Active StackOverflow askers distribute their work in a less egalitarian way (i.e., focus their attention more) than developers that do not ask questions.

Interaction between activities

- Time series defining evaluation latency and the response latency in this context (github and stackoverflow)

Results

- For active committers, asking questions on StackOverflow catalyses committing on GitHub. Similarly, for active committers, answering questions on StackOverflow catalyses committing on GitHub.
- For the most active answerers as well as for developers that do not answer any questions at all, their StackOver- flow activities accelerate their GitHub committing.

Conclusions and Future Work