International Journal of Open Source Software and Processes
Volume 11 • Issue 4 • October-December 2020

Open Source Software Development Challenges:

A Systematic Literature Review on GitHub

Abdulkadir Seker, Sivas Cumhuriyet University, Turkey

Banu Diri, Yıldız Technical University, Turkey

Halil Arslan, Sivas Cumhuriyet University, Turkey

Mehmet Fatih Amasyalı, Yıldız Technical University, Turkey

ABSTRACT

GitHub is the most common code hosting and repository service for open-source software (OSS) projects. Thanks to the great variety of features, researchers benefit from GitHub to solve a wide range of OSS development challenges. In this context, the authors thought that was important to conduct a literature review on studies that used GitHub data. To reach these studies, they conducted this literature review based on a GitHub dataset source study instead of a keyword-based search in digital libraries. Since GHTorrent is the most widely known GitHub dataset according to the literature, they considered the studies that cite this dataset for the systematic literature review. In this study, they reviewed the selected 172 studies according to some criteria that used the dataset as a data source. They classified them within the scope of OSS development challenges thanks to the information they extract from the metadata of studies. They put forward some issues about the dataset and they offered the focused and attention-grabbing fields and open challenges that we encourage the researchers to study on them.

KEYWORDS

GHTorrent, GitHub, Open-Source, OSS, SLR, Systematic Literature Review

INTRODUCTION

Thanks to distributed version control systems such as Git, Mercurial, etc., open-source development platforms have reached a considerable number of users. The most common of these platforms is GitHub (based on git). GitHub has become the world's largest code server with more than 40 million developers hosting and collaborating over 100 million repositories.

On platforms such as GitHub, the development process is distributed. Developers can participate in a project, contribute, discuss bugs with each other, and write comments about code from various locations. In this way, a considerable amount of textual, numerical and network or collaboration-based features about the projects and developers are extracted from the platform. Besides, GitHub includes many social relations among users or projects. GitHub is the most common code hosting and repository service for open-source software projects. For the researchers that focus on software engineering, the content of this platform provides many valuable sources. Most of the studies about this domain use GitHub as a data source because of easy to access, amount of data, and diversity of features. In this context, we think that is important to conduct a literature review on studies that used GitHub data.

DOI: 10.4018/IJOSSP.2020100101

There are several options to reach GitHub data. In a survey study which is given the usage rates of GitHub dataset, they addressed that the most used dataset is GHTorrent (34%) in the articles that are reviewed according to the certain criteria (Cosentino, Luis, & Cabot, 2016). In Cosentino's systematic mapping study, the GHTorrent dataset is in the lead with a 41\% use rate (Badashian, Shah, & Stroulia, 2015; Cosentino, Canovas Izquierdo, & Cabot, 2017). In the another study, GHTorrent is the most cited dataset (Kotti & Spinellis, 2019). The GHTorrent dataset was developed by Georgios Gousios in the software engineering department at Delft University of Technology(Gousios, 2013). The dataset is generated by systematically crawling with the GitHub API and includes information about all public projects and users on the platform. GHTorrent stores some information about repositories, projects, issue descriptions, comments, and pull request (PR) conversations in 26 relational tables totally.

We saw from other systematic literature review (SLR) papers that some studies can be missed when reviewing with a text-based (keyword) search from search engines or digital libraries. Because of that, to reach the studies, we conducted this literature review based on a GitHub dataset source study instead of a keyword-based search in digital libraries. Due to GHTorrent is the most widely known and used GitHub dataset according to the literature, we considered the studies which cite this dataset for the systematic literature review.

In this study, we offered to find out the topics of all studies and classified them. We focused on the studies with the context of open-source software development. We divided the studies into some categories and challenges. Besides, some distributions (type, venue, year, method, data, topic) have been obtained from the studies that used the dataset. We show which challenges are mentioned in the studies and how each study is using the dataset. Thus, we hope the study guided the researchers who interest in software engineering challenges with open-source systems. We formed this review following these research questions:

RQ1: What are the trends of open-source software development challenges?
RQ2: What are the handicaps/cons of GHTorrent?
RQ3: What are the open challenges that have not yet been studied with this dataset?

In this context, we reviewed the articles which use GHTorrent and offered a systematic mapping study. We applied 3 phased systematic literature review protocol as suggested by Kitchenham (Brereton, Kitchenham, Budgen, Turner, & Khalil, 2007). Firstly, we developed a review method using citations of the main paper of the dataset. Then, we conducted a review as extract trend topics from metadata of studies and made assessments. Finally, we revealed some discussions and open challenges. The protocol and details are given Figure 1. We used a cross-checked mechanism (two of the authors) while finding studies and classifying them.

METHODOLOGY

Developing Review

In these other SLR studies, they noticed that some studies can be missed when reviewing with a keyword based search from digital libraries (Khan & Keung, 2016; Schreiber & Zylka, 2020). Because of that, we followed the citation of the main study of the dataset. We used an application[1] to extract all citations of the GHtorrent's study. All 332[2] studies which cited the main study of GHTorrent (Gousios, 2013) were reviewed. We applied exclusion criteria similar to the recently published an SLR study (Schreiber & Zylka, 2020). We exclude the studies that were written in any language other than English, paid studies, and reports/books/theses (Table 1). In addition, the articles refer GHTorrent only as related works or similar dataset were also eliminated.

After we applied the exclusion protocol, we reviewed 172 studies. 49 of the studies were published in journals, and the remaining 123 were published in conferences.

Related Content