Prediction Models and Techniques for Open Source Software Projects: A Systematic Literature Review

M.M. Mahbubul Syeed, Tampere University of Technology, Tampere, Finland
Imed Hammouda, Chalmers and University of Gothenburg, Göteborg, Sweden
Tarja Systä, Tampere University of Technology, Tampere, Finland

ABSTRACT

Open Source Software (OSS) is currently a widely adopted approach to developing and distributing software. For effective adoption of OSS, fundamental knowledge of project development is needed. This often calls for reliable prediction models to simulate project evolution and to envision project future. These models provide help in supporting preventive maintenance and building quality software. This paper reports on a systematic literature survey aimed at the identification and structuring of research that offer prediction models and techniques in analyzing OSS projects. In this review, we systematically selected and reviewed 52 peer reviewed articles that were published between January, 2000 and March, 2013. The study outcome provides insight in what constitutes the main contributions of the field, identifies gaps and opportunities, and distills several important future research directions.

Keywords: Fault Prediction, Open Source Software, OSS Community, Prediction, Systematic Literature Review

1. INTRODUCTION

The use of Open Source Software (OSS) is increasingly becoming part of the development strategy and business portfolio of more and more IT organizations. This is, for example, demonstrated by the growing numbers of downloads of OSS code by companies (Samoladas; Angelis; & Stamelos, 2010). The primary motivation is that OSS can offer huge benefits to an organization, with...
minimal development costs while taking advantage of free access to code and high quality levels driven by the power of distributed peer review (Capiluppi & Adams, Reassessing brooks law for the free software community, 2009). Successful OSS projects such as Eclipse have reached thousands of downloads per day (Eclipse, 2013). However, such projects are typically complex, both from the point of view of the code base and with respect to the developer and user community. They may consist of a wide range of components, and come with a large number of versions reflecting their development and evolution history.

In order to adopt an OSS component effectively, an organization often needs fundamental knowledge of the project development, composition, and the possible risks associated with its use. This is because OSS code is primarily developed outside the company by an ultra-wide distributed community (Thy, Ferenc & Siket, 2005; Samoladas, Angelis, & Stamelos, 2010). In particular, organizations might need to understand how an OSS project may evolve, as this may impact the future of the organization itself. From a proactive perspective, foreseeing the evolution of an OSS component may provide the organization with useful information including the kind of maintenance practices, resources, and strategic decisions need to be allocated and adopted in supporting their development strategies.

Accordingly, a wide range of prediction models have been proposed by the research community for the purpose of simulating the evolution and approximating the future of OSS projects, with regard to various aspects. For instance, a number of methods supporting error prediction have been developed to provide valuable information for preventive maintenance, and for building quality software. An example prediction scenario has been to foresee potential error prone segments of the code base for tracing down the modules that would most likely require future maintenance tasks (Thy, Ferenc, & Siket, 2005; Yuming & Baowen, 2008). Despite the variety and volume of OSS prediction studies, it has been argued that the efforts for analyzing the evolutionary behavior of OSS systems still lag behind the high adoption levels of OSS. Furthermore, the focus of OSS prediction studies in general has been restricted to a small number of projects, which limits the generalizability of the methods and results. Such claims thus need empirical evidence (Russo, Mulazzani, Russo, & Steff, 2011).

In this paper, we present a literature review with the aim to provide an in-depth analysis of the prediction research work targeted at analyzing open source projects. The literature that explores both the technical and social dimensions of the OSS projects for constructing prediction models is studied. Our overall goal is twofold. First, we offer a single point reference to the state-of-the-art study on the topic, and second, we distill gaps and opportunities as future research directions.

To carry out this review we developed a review protocol following the guidelines presented in (Kitchenham,