White Paper for Categorizational Study on Characteristics of Developer Behaviors

Jincheng He, 6792671766, jinchenh@usc.edu

March 1, 2021

Abstract

Nowadays, it is common to develop

1. Introduction

This direction mainly focuses on investigating the developer behaviors including how they contribute to the software project, especially to open source software where we can get enough metadata for analysis to evaluate how different behaviors impact the software quality and how to improve this process for the sake of better software quality. Specifically, this research will focus on modelling the change pattern in code when contributors make the commits and how the code impacts the quality metrics. In this direction, research has been conducted to investigate when, where, how and what the developers contribute to the projects but there is still space for research from the coding side. To reveal details, for example, the change types and contents, in code level can reveal further detail how different changes related to developer behaviors and how they impact the software quality, which is represented in the software quality metrics.

This direction could succeed since the current techniques in machine learning, statistics, natural language processing will be sufficient to support this research. Once it succeeds, we can provide further instructions in coding, if possible, more reliable coding standards which can lead to an improvement and standardization of coding in the software engineering area.

This direction may take more than ten years since it could be a gradual improvement. The success of this direction depends on how the applied techniques evolve in the coming years. The midterm milestone could be a

reasonable high prediction accuracy while the final exam milestone could be the completion of the new systematic coding standards.

Research has been done to investigate the impact of some behaviors of developers when they contribute to software, especially open source software. However, in the previous research in this area, they haven't reveal the correlation between the change type and the code as well as their impact on the software quality. In this research, we start with categorization of commit changes in open source software repositories and show their impact on the software quality, collected by using different static analysis tools, by which we get software metrics for each revision of the software. In the end of this paper, we propose a new categorization for the commit change, based on reading the code and the commit message. We also apply neural network to train a model to predict the potential software quality changes after a certain kind of change with an accuracy of 90

In previous research, we have been proven there exist correlation between the change type and the software quality, which is represented by the software quality metrics. However, this is not enough to put this research into valuable application. In this research, we refine the categorization by applying a natural language processing approach to the code to parse it. In this way, we can model the change into a lower level, and make more accurate prediction of what is happening in the repository of the open source software. The results shows we can predict the change of software quality change based on the code with a high accuracy of 95

In previous research, the researchers use natural language processing methods to predict the possible changes to the software quality based on the code changes. This means we can provide guidelines to some extent for software contributors when they commit to a software. However, the previous study mainly focus on training the model with current data which has its limitation. In this research, we collect new data from different software repositories with more than 100 coding styles, including most of existing coding styles. By comparing the different coding styles and training the new prediction model, we come up with a guideline for how to code more efficiently with less software quality issues. The results turn out to be valuable, providing a 95