

End of Quarter Report

Accomplishments

This quarter I have successfully begun my work towards my thesis. This was my first quarter enrolled in the thesis courses. I began this quarter working on the original thesis, Automated Classification of Ticket Types. The goal of this thesis was to create a model to automatically classify JIRA and Bugzilla tickets using various metrics including title, description, dates, severity, and comments on the issue. I read research papers that discussed the importance of classifying the tickets correctly and how to predict an issue as a bug or not. I wrote two programs to calculate metrics for JIRA and Bugzilla issues. These can be found in the Code folder under the Code for Previous Thesis folder. After discovering a classification model has already been done, the thesis changed.

The current thesis topic focuses on Affected Versions (AV) in JIRA bug reports. We wish to predict the AV accurately and test how the affected versions increases bug prediction accuracy. I began this thesis by reading research papers that focus on why the quality of data is vital in defect predictions and models that attempted to predict the origin of a bug. Next, I began research into the quality of AV in the issue reports of Apache JIRA projects. I created a few programs to get more insight regarding the Apache JIRA issue reports and the AV recorded. I first started off by writing a program to get the percentages of recorded AV. Next, I wrote a program to calculate the release of the creation date, release of the fix date, and the list of AV for each closed bug in an Apache JIRA project. Here, I noticed data was not accurate in that some AV did not have release dates. So, I wrote a program to find out what releases were missing release dates for each project. These three programs can be found in the Progress Towards RQ2 in the Code folder. After noticing a few more issues with the AV recorded, I wrote a Quality Model which calculated the number and percent of bugs with each type of AV problem for each project. We defined a new research question, RQ0, which inspects the quality of the recorded AV. This quality model can be found in the RQ0 code in the Code Folder.

Challenges

As mentioned previously, various challenges came along during the start of my thesis. The first challenge was discovering a model had already been created for the topic of my first thesis. Once changing and redefining my thesis, I found the data for AV was not as accurately recorded as desired. Various issues in the AV existed such as missing release dates, AV recorded after the fix release, and arbitrary AV that are not actual releases but terms such as all. We altered the thesis to investigate into the quality of AV to account for these problems.

Goals

Next, I hope to expand on how the quality of AV is relevant and if it affects defect predictions. I hope to determine whether these problems in the AV are fixable problems or whether they seriously affect defect predictions when present. I hope to finish all research questions by the end of next quarter, so I can spend the remaining of summer writing my thesis and defending.

What I Have Learned

I have learned this quarter that finding quality data in software engineering is not only vital, but a very difficult task. Issue reports are not always filed and recorded with the best accuracy which affects metrics that contribute to defect predictions. Making accurate predictions is important to save time, effort, and cost in the industry of software engineering.