Author: Xin Ye, email: xy348709@ohio.edu

August 29, 2014

Introduction

This is the dataset that was used to evaluate the learning-to-rank approach described in the FSE 2014 paper "Learning to Rank Relevant Files for Bug Reports Using Domain Knowledge". The dataset contains bug reports and the corresponding commit history for six open source Java projects: AspectJ, Birt, Eclipse Platform UI, JDT, SWT, and Tomcat.

How to obtain the dataset

Download the dataset.zip archive from the link below:

- http://figshare.com/articles/The_dataset_of_six_open_source_Java_projects/951967

Once unzipped, the dataset folder will contain the dataset in Excel format (xls) and XML (xml) format. The folder also contains a database file "dataset_fse2014.sql" and the corresponding "README.MD" file.

Dataset structure

File list:

1. AspectJ.[xls|xml] – The bug reports and commit history of AspectJ.
2. Birt.[xls|xml] – The bug reports and commit history of Birt.
3. Eclipse_Platform_UI.[xls|xml] – The bug reports and commit history of Eclipse Platform UI.
4. JDT.[xls|xml] – The bug reports and commit history of JDT.
5. SWT.[xls|xml] – The bug reports and commit history of SWT.
6. Tomcat.[xls|xml] – The bug reports and commit history of Tomcat.

XLS/XML Headings:

- *bug_id* -- refers to the bug report id.
- *summary* -- refers to the bug report summary.
- *description* -- refers to the bug report description.
- *report_time* -- refers to the bug report report time.
- *report_timestamp* -- refers to the bug report report timestamp.
- *status* -- refers to the status of the bug report.
- *commit* -- refers to the SHA-1 hash id for the commit that fixed the bug report.
- *commit_timestamp* -- refers to the commit timestamp.
- *files* -- contains the full path of every Java file that was fixed in this commit.

- *result* -- contains the position of every positive instance in our ranked list result.

How to obtain the source code

1. **AspectJ**: *git clone* git://git.eclipse.org/gitroot/aspectj/org.aspectj.git
2. **Birt:** *git clone* https://git.eclipse.org/r/p/birt/org.eclipse.birt
3. **Eclipse Platform UI:** *git clone* https://git.eclipse.org/r/p/platform/eclipse.platform.ui
4. **JDT:** *git clone* https://git.eclipse.org/r/p/jdt/eclipse.jdt.ui
5. **SWT:** *git clone* https://git.eclipse.org/r/p/platform/eclipse.platform.swt
6. **Tomcat:** *git clone* git://git.apache.org/tomcat.git

A before-fix version of the source code package needs to be checked out for each bug report. Taking Eclipse Bug 420972 for example, this bug was fixed at commit 657bd90. To check out the before-fix version 2143203 of the source code package, use the command *git checkout 657bd90~1*.

Efficient indexing of the code

If bug 420972 is the first bug processed by the system, we check out its before-fix version 2143203 and index all the corresponding source files. To process another bug report 423588, we need to check out its before-fix version 602d549 of the source code package. For efficiency reasons, we do not need to index all the source files again. Instead, we index only the changed files, i.e., files that were “Added”, “Modified”, or “Deleted” between the two bug reports. The changed files can be obtained as follows:

- **Added**: *git diff --name-status 2143203 602d549 | grep ".java$" | grep "^A"*
- **Modified**: *git diff --name-status 2143203 602d549 | grep ".java$" | grep "^M"*
- **Deleted**: *git diff --name-status 2143203 602d549 | grep ".java$" | grep "^D"*