Tracking Line Changes in Source Code Repositories

Francesca Arcelli Fontana
University of Milano-Bicocca
Viale Sarca, 336 Milano, Italy
arcelli@disco.unimib.it

Marco Zanoni
University of Milano-Bicocca
Viale Sarca, 336 Milano, Italy
marco.zanoni@disco.unimib.it

Categories and Subject Descriptors

D.2 [**Software**]: Software Engineering; D.2.7 [**Software Engineering**]: Distribution, Maintenance, and Enhancement—*Version control*

General Terms

Measurement, Experimentation

Keywords

Repository analysis, line changes, file changes, correlation

1. CONTEXT

Previous research determined that the analysis of file changes in software repositories is useful for maintenance activities, like defect prediction. Changes rarely modify the entire file contents, but are usually localized in specific code regions.

2. GOAL

We aim to investigate the effects of tracking changes to single lines of code, opposed to changes applied to entire files. We apply a line change tracking algorithm for measuring changes occurred to the single lines of code of each file of eight analyzed code repositories. With these data, we try to answer to the following research questions:

RQ1 Does the number of file and line changes provide the same information?

RQ2 Does the usage of the number of changes per file or line, provide different rankings of the same files?

3. METHOD

We defined a technique to track changes made to the lines of each file in a repository. To track changes, the algorithm analyzes unified diff format patches between files consecutive versions. For each line, a descriptor tracks the list of changes made to the line, recording the timestamp, the version identifier, the author, and the position of the changed line. Edits are represented by the deletion of lines followed by the addition of the former lines, incorporating the change. Changes are assigned to lines following simple rules, based on the position of lines in change blocks. We computed and recorded the number of changes for all files and their lines. We then aggregated this data on files using different statistics, i.e., minimum, maximum and median. Line and file statistics have been compared using Kendall τ correlation.

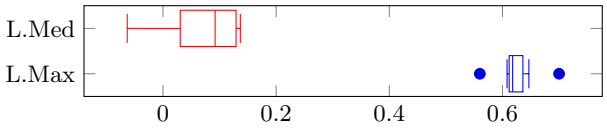

Figure 1: Num. file and line changes (correlation)

4. RESULTS

Figure 1 shows the correlation between the number of file changes and the maximum and median number of line changes for the respective file. We did not find total correlation between the measures. To answer RQ1, the number of changes in files and lines provide similar, but different information. In particular, maximum line changes are not totally correlated with file changes. As for RQ2, we compared the ranking produced by the two different approaches, and we found very different positioning of certain files. For example, build files tend to have very high rankings when considering file changes, but not when considering line changes.

5. CONCLUSIONS

Line change tracking could be helpful in the same areas where file change measures have been applied. In many cases, line changes can be a more precise indicator than file changes. We are replicating this study using time between changes. All analyses are supported by our VCS-Analyzer (`http://essere.disco.unimib.it/reverse/VCSAnalyzer.html`) tool.

ESEM '14, September 18–19, 2014, Torino, Italy.

http://dx.doi.org/10.1145/2652524.2652597