Aman Sharma

Martin Monperrus

Computer Science

02 November 2020

Disagreement between Myers and Histogram

This report outlines the disagreement between the two `git diff` algorithms - Myers and Histogram on the basis of:

1. The difference in NLA & NLD
2. The difference in location of the changed lines

The source code of the implementation can be found here[1].

Task 1: Difference in number of commits based on NLA & NLD

Observations

	Myers	Histogram
NLA	33289	33544
NLD	47740	47998

Total No. of Commits: 836
No. of different commits: 36
%Difference: 4.30622009569378

[1] https://github.com/algomaster99/myers-vs-histogram/blob/master/difference_in_commits.ipynb

Conclusions

1. The no. lines modified are greater when the diff output is generated using *Histogram* as the diff algorithm. The studies concerning the size of commits[2] will vary with *Histogram* algorithm producing greater sizes of commits. This can also impact the contributions of individual contributors.
2. The larger number of deleted lines in the case of *Histogram* algorithm is equivalent to more "valid bug-related lines" as the SZZ algorithm will result in a higher number of candidate of bug-related lines. It is true that the algorithm tags a lot of unsuspicious lines as "valid" too but overall, it is unlikely that a potential candidate of bug related line will be missed out. Meanwhile, there is a possibility that using *Myers* some of the suspicious deleted lines are ignored due to their lesser number.
3. Code coverage software can run over more number of modified lines and eventually give a better coverage report in the case of *Histogram*'s result.
4. The slight variation (+2 commits) in "%Difference" can be attributed to the following reason:
    a. Overwriting of git history

[2] The size of a commit, or commit size, is defined as the sum of the number of source code lines (or source lines of code, SLoC) added, removed, or changed. Ref: https://flosshub.org/sites/flosshub.org/files/Estimating%20Commit%20Sizes%20Effciently.pdf

Task 2: Difference in number of commits based on the location of changes

Observations

Sum of	Myers	Histogram
Contiguous sequences of added lines	20252	20423
Contiguous sequences of deleted lines	10139	10252

Total No. of Commits: 836
No. of different commits: 10
%Difference: 1.1961722488038278

Conclusions

1. The larger code hunks obtained using *Histogram* algorithm ensures more readability of differences. The larger hunks show the changed lines more systematically and thus, making it easier to analyze the changes.
2. Variation in the location of modified lines can lead to misidentification of bug introducing lines using SZZ algorithm. The question which is better to identify these lines correctly needs more analysis to be answered.
3. The huge variation in the output of difference in commits is due to an assumption I have taken. I assumed the following three conditions to find out the number of commits with a difference in locations of modified lines:
    a. Number of added lines using *Myers* == Number of added lines using *Histogram*

b. Number of deleted lines using *Myers* == Number of deleted lines using *Histogram*
c. If two similar lines exist, their position must not be equal.

This ensured that the cases where NLA and NLD are unequal do not mix with the current case. However, the paper suggests, “If the positions of each changed line of code were the same, we considered the results the same; otherwise, the results were considered different. ” to identify the different locations of modified lines. This was unclear to me as the changed line could be not the same because the line being compared won’t exist in the other output due to differences in the number of lines.