DEREK.MA | DEREK.MA@CONNECT.POLYU.HK

COMP3211 Homework Three

MA Mingyu, Derek - 14110562D

November 5, 2017

1 Paper Summary

Copying and pasting code, a very common practice in software development, is prone to introducing bugs especially for large projects like operating systems. A new tool called *CP-Miner* is introduced in this paper, which is a scalable tool to detect copy-paste (CP) code segments and related bugs especially the forget-to-change bugs using data mining techniques.

Compared to existing similar tools like *Moss*, *CCFinder* and so on, CP-Miner is scalable and is capable of CP-related bugs detection. It can tolerate statement insertions and modifications in CP as well.

For detecting CP code, CP-Miner converts the problem into a frequent subsequence mining problem and then follows the steps: parsing source code, mining basic CP segments, pruning false positives and finally composing larger CP segments. To detect CP related bugs, *UnchangedRatio* is defined as the ratio of unchanged occurrences and lower *UnchangedRatio* indicates the appearance of bugs.

The evaluation results show that CP-Miner can detect more CP segments than CCFinder, and it can also find many CP related bugs in Linux and FreeBSD. Some statistical patterns on CP characteristics in large software are presented in this paper as well.

References

[1] Li, Z., Lu, S., Myagmar, S., Zhou, Y. (2004, December). CP-Miner: A Tool for Finding Copy-paste and Related Bugs in Operating System Code. In OSdi (Vol. 4, No. 19, pp. 289-302).