Exact and Near-miss Clone Detection in Spreadsheets

Felienne Hermans, Delft University of Technology

ABSTRACT

Spreadsheets are used extensively in business, in many domains. The applicability of software engineering methods to spreadsheets has been a topic of research for several years [2], but the main focus has been on analyzing the formulas, and not on analyzing the data in the spreadsheets. One of the factors that plays a role in spreadsheet data quality is the occurrence of clones in the spreadsheet data.

Clones in data are caused by copy-pasting. This is a very common practice in spreadsheet use, however, it can have a negative impact on the spreadsheet's quality, since 1) editing the copied data needs to be done in multiple places increasing maintenance effort and 2) when editing, some copies might be forgotten, leading to errors.

Clone detection has been proven useful in the realm of source code analysis [3], in two different forms: exact clones, and clones that differ slightly, called near-miss clones [1]. Because of the success of clone detection and removal in source code, it seems feasible to research the applicability of both techniques on clones in spreadsheet data. Our work shows that this is a promising avenue.

BODY

Clone detection in spreadsheets is useful both to reveal opportunities for improving the spreadsheet and to detect actual errors.

REFERENCES

[1] J. R. Cordy, T. R. Dean, and N. Synytskyy. Practical language-independent detection of near-miss clones. In H. Lutfiyya, J. Singer, and D. A. Stewart, editors, *CASCON*, pages 1–12. IBM, 2004.

[2] M. Erwig. Software engineering for spreadsheets. *IEEE Software*, 26(5):25–30, Sept. 2009.

[3] R. Fanta and V. Rajlich. Removing clones from the code. *Journal of Software Maintenance*, 11(4):223–243, 1999.