Abstract

This paper presents a new technique for clone detection using sequential pattern mining titled EgyCD. Over the last decade many techniques and tools for software clone detection have been proposed such as textual approaches, lexical approaches, syntactic approaches, semantic approaches …, etc. In this paper, we explore the potential of data mining techniques in clone detection. In particular, we developed a clone detection technique based on sequential pattern mining (SPM). The source code is treated as a sequence of transactions processed by the SPM algorithm to find frequent itemsets. We run three experiments to discover code clones of Type I, Type II and Type III and for plagiarism detection. We compared the results with other established code clone detectors. Our technique discovers all code clones in the source code and hence it is slower than the compared code clone detectors since they discover few code clones compared with EgyCD.

References
15. Chao Liu, Chen Chen, Jiawei Han and Philip S. Yu. GPLAG: Detection of Software Plagiarism by Program Dependence Graph Analysis. In the Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’06), pp. 872-881, Philadelphia, USA, August 2006.
17. A. Leitlao, Detection of Redundant Code Using R2D2, Software Quality Journal,
26. Hamid Abdul Basit, Member, IEEE, and Stan Jarzabek, A Data Mining Approach for Detecting Higher-level Clones in Software.

Index Terms
Computer Science Information Sciences

Keywords
Sequential Pattern Mining, Clone Detection, Data Mining