Abstract

This research presents a parallel and distributed data mining approach to code clone detection. It aims to prove the value and importance of deploying parallel and distributed computing for real-time large scale code clone detection. It is implemented this approach in a family of clone detectors, called PD EgyCD (Parallel and Distributed Egypt Clone Detector). In this approach, This research builds on an earlier work of the authors for code clone and plagiarism detection using sequential pattern mining by adding parallelism and distribution to our earlier tool EgyCD. Our approach uses data mining through a tailored Apriori-based algorithm for code clone detection. And it uses parallelization and distribution to achieve excellent performance to scale up to clone detection on very large systems. This approach has been implemented as a database application which leverages the capabilities of modern database tools. Two versions have been developed of this distributed technique. The first one uses client-server technique in which all clients and the server deal with only one database. The second one uses agents where each client acts as a separate agent and has its own database and after working on a sub-problem, it submits its partial solution to the server to finally get the complete solution (set of code clones). Experiments show that agents technique is faster than client-server one.
Distribution enhances performance very much. Speed improvement is a function of the number of clients/agents used. Our conclusion is that data mining, combined with parallel and distributed computing, can efficiently be deployed for code clone detection of very large systems.

References


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.


S. Livieri, Y. Higo, M. Matushita, K. Inoue, Very-Large Scale Code Clone Analysis and Visualization of Open Source Programs Using Distributed CCFinder: D-CCFinder, Graduate School of Information Science and Technology, Osaka University 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan, 2007


Jiawei Han, Micheline Kamber: Data Mining – Concepts and Techniques, Kaufmann, 2001.

Index Terms

Computer Science Information Sciences

Keywords

Code clones textual approach lexical approach syntactic approach clone types parallel code clone detector
distributed code clone detector

clone relation terminologies

data mining

apriori property

sequential pattern mining