API Analysis using Association Rule Mining

M. Monperrus - M2 IAGL - 19 novembre 2013

Goal of the exercise:
- understand the concept of data-mining for software engineering
- practice with the apriori algorithm

Context:
Software APIs are always used in certain ways. Those regularities can be captured by frequent itemsets and association rules. Those rules complement the existing documentation and can even be used to find bugs.

Reference:

Data Mining Library Reuse Patterns using Generalized Association Rules (Amir Michail), 2000.

Getting Started:
- Read the introduction of of the paper
- Run Borgelt's apriori[1] on a toy file (for instance Label.dat on Moodle)
- Download and understand vmlinux-debug-2.6.24-29-386.asm.bz2, the extraction[2] of all function calls of the Linux kernel.

Tasks:
Collect data: already done, see vmlinux-debug-2.6.24-29-386.asm

Pre-process data: write a small program that transforms the file produced by objdump in order to create a file that is compatible with apriori. One line is a function, each item of of aline is a function call.

Collect association rules: run apriori on the curated data to collect association rules.

Understand the rules: choose 2 common rules and understand it. This means browsing the related code, the documentation and the web to fully grasp what it means.

Report:
A 2 pages PDF report explaining the rules. For each rule, give:
- the support and confidence
- one code snippet illustrating the rule
- the explanation of the rules in one paragraph
- a link to the most related documentation

Presentation topics: section 3 and 4/5 of the paper

1 http://www.borgelt.net/apriori.html
2 extracted with objdump -D vmlinux-debug-2.6.24-29-386 | egrep "call |Diss|(:$)"