Large scale configurable text matching for log change and anomaly detection

Sandvine (ex Procera Networks) develops equipment used by major network operators to ensure high-quality connectivity for more than 1.7 billion users.

The equipment generates a considerable amount of log data in various situations. This thesis work concerns the analysis of this log data to improve understanding that can be derived from these logs. In particular, two objectives are of interest:

1) Locating changes in equipment boot log messages. As a system boots up, it generates a large amount of log data. During system development for these systems it is very relevant to know if the boot-up log messages change when changes are made in the codebase. However, since the logs are large it is not feasible for a human to go through the logs and look for changes. It is also not possible to make a simple diff, as there are natural variation in a subset of the log message that should not be flagged up to the developers. This part of the thesis work concerns building a system that can be continuously configured, i.e “learnt”, by developers to detect and show relevant log message changes, but give no indication for log messages known to be harmless or only varying showing natural variation.

2) Locating changes/anomalies in operational log data. When deployed, the equipment continuously generates operational log data that reflects the processing load, memory usage etc. This task is concerned with detecting changes in these logs over time, with the aim to detect when a long-time change has occurred in the monitored metrics. Such detected metric changes can then be compared to when firmware updates were applied, configuration changes were made, etc. This will allow a deeper understanding of any interaction between firmware/configuration changes and operational resource usage.

The solution for these tasks is expected to employ a number of techniques such as hashing, approximate text matching/regular expressions, and the use efficient algorithms to perform these. Which task to focus on initially, and the solution approach will be jointly decided in discussion between the student and Sandvine.

This thesis work is done with Sandvine, but much of the work will be done in Karlstad with a couple of visits to company premises in Varberg and Malmö. The work is expected to be done using Python and associated libraries. Prior Python experience is not a requirement, but good general programming proficiency and willingness to learn Python is important. This thesis work is the continuation of previous very successful thesis cooperation between Sandvine and KAU, where four previous KAU CivIng students already have successfully completed thesis work.

The technical work part of the thesis will be jointly supervised, with a Sandvine senior technical engineer and in close cooperation with KAU researchers active in the area.

The writing part of the thesis will be supervised by the KAU supervisor (i.e one of the active researchers).

Monetary compensation will be provided from Sandvine for students participating in this thesis project.

For questions or more details, please contact Johan Garcia at johan.garcia@kau.se, or visit me at room 21F426