Large scale configurable text matching for change and anomaly detection in log files

Sandvine 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 is can be continuously configured, i.e. “learnt”, by developers to detects 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.

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