Kerf: machine learning to aid intrusion analysts

Javed Aslam*, Sergey Bratus†, David Kotz†, Ron Peterson†, Daniela Rus‡

*Northeastern University, jaa@ccs.neu.edu

†Department of Computer Science and Institute for Security Technology Studies,
Dartmouth College, {sergey, dfk, rapjr}@cs.dartmouth.edu

‡Massachusetts Institute of Technology, rus@csail.mit.edu

Kerf is a toolkit for post-hoc intrusion analysis of available system logs and some types of network logs. It takes the view that this process is inherently *interactive* and *iterative*: the human analyst browses the log data for apparent anomalies, and tests and revises his hypothesis of what happened. The hypothesis is alternately *refined*, as information that partially confirms the hypothesis is discovered, and *expanded*, as the analyst tries new avenues that broaden the investigation.

Since log analysis is a repetitive and laborious task, Kerf's tools help automate this process with techniques drawn from the Machine Learning community. We believe that this approach will lead to a new breed of more efficient log analysis tools. In particular, we have applied it to the following tasks.

- **Parsing logs..** Data for analysis often comes from freetext records in many varying formats (e.g., UNIX syslog) and must be parsed out.

  We use simple Natural Language Processing and Machine Learning techniques to help the administrator bootstrap a set of patterns for parsing freetext logs (e.g., for future statistical analysis or DB storage). Our tool helps the user to start with large and diverse syslogs, and rapidly generate patterns (such as "`Accepted %auth_method for %user from %ip port %port`") for various kinds of records, instead of resorting to `awk`, `grep`, command-line perl, and so forth.

- **Data organization..** Once parsed, large sets of records are usually grouped, sorted and summarized as a first step of analysis.

  We use classification and clustering algorithms to choose the best way of presenting and summarizing the data. In particular, record sets are presented in dynamic tree form, to make statistical facts about distribution and mutual dependencies of values apparent and anomalies more prominent. The tree views are built adaptively, based on entropy of values in the log records within a result set and their mutual information statistics. With our data organization algorithms, the user can try new ways of browsing the logs, e.g., by record similarity (adaptively, within the current set, or based on previous markup), or by correlation and co-occurrence of record's fields.

- **Describing and implementing correlation..** Writing queries for correlation of records on their time and values contained is laborious and error-prone even if the values are parsed out and held in a SQL database.

  We use a SQL-based domain-specific language, SawQL, for expressing relative time and value correlations more concisely and naturally. Each SawQL query describes a sequence of related events. Corresponding SQL queries are then formed automatically, and their results are fed to data organization algorithms.

- **Interactive hypothesis generation..** Important characteristics of an interesting set of log records may not be immediately obvious from summarization or correlation alone.

  As the administrator browses logs and marks suspicious records in them, a concise description of the marked set (a SawQL query that distinguishes the marked subset from the rest of the log) is desired but not easy to generate manually. We are developing a semi-automated learning solution based on the Minimal Description Length approach for generating queries, which can be thought of as descriptions of intrusion traces.

The Powerpoint slides from the accompanying presentation are available at: http://kerf.cs.dartmouth.edu/slides/usenix-sec04.pdf

This project was supported under Award No. 2000-DT-CX-K001 from the Office for Domestic Preparedness, U.S. Department of Homeland Security. Points of view in this document are those of the author(s) and do not necessarily represent the official position of the U.S. Department of Homeland Security.