Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

Automatisierte semantische Anreicherung von Logdaten

Diplomand
Timon Kopp Ronchetti

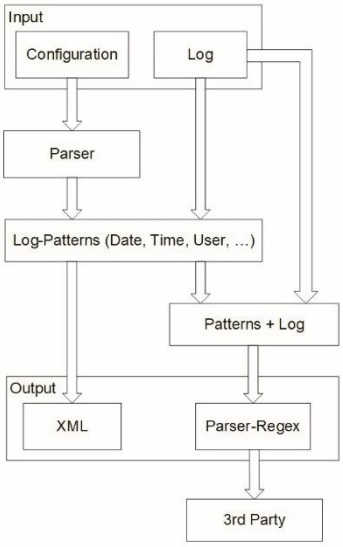
Dozent
Karl Rege

Durch die immer dichtere Vernetzung von elektronischen Geräten wächst zwangsläufig auch die Menge der Events im Netzwerk. Die hohe Diversität der Hersteller, Technologien und Einsatzbereiche resultieren in einer nahezu unüberschaubaren Ansammlung von Informationen aller Art. Um dennoch eine ganzheitliche Sicht über die gesammelten Events zu erhalten, werden SIEM-Systeme eingesetzt, zum Beispiel für das Erkennen von Trends, Mustern und das Hervorheben relevanter Informationen. Dafür müssen die eingelesenen Events vorab in ein einheitliches Log-Format übersetzt werden. Da sich bisher noch kein offizieller Standard für Protokollnachrichten etabliert hat, unterscheiden sich die Logs in Format, Struktur und Reihenfolge, weshalb sich eine Normalisierung als sehr problematisch erweist.

Diese Arbeit beschäftigt sich damit, die Bestandteile einer Logdatei automatisch auf ihre semantische Bedeutung zu identifizieren und diese als Regular Expressions auszudrücken. Die Ausgabe soll darüber hinaus so konzipiert sein, dass sie von externen Systemen für Parser und Filter verwendet werden kann.

Um dieses Ziel zu erreichen, wurden die Eigenschaften und Merkmale diverser Logdateien und Standards gesammelt, analysiert und zusammengefasst. Die daraus erhaltenen Informationen resultierten in einer Kombination aus Regular Expressions und Analyse-Regeln, welche auf die einzelnen Protokolleinträge einer Logdatei angewendet werden, um so die gesuchten Bestandteile zu identifizieren. Nach erfolgreicher Lokalisierung der Elemente wird geprüft, welche der erfassten Patterns für die Identifizierung verwendet wurden. Diese werden anschliessend zu einigen wenigen Regular Expressions zusammengefasst, mit welchen sich alle Zeilen ausdrücken lassen.

Die Resultate werden anhand der Anzahl identifizierter Bausteine, ihrer semantischen Korrektheit und der praktischen Verwendbarkeit der Ausgabe gewertet. Es zeigt sich, dass die wichtigsten Elemente einer Logdatei, wie zum Beispiel Datum, Uhrzeit, Severity, Server und Benutzer, grösstenteils korrekt identifiziert und beschriftet werden. Defizite kristallisieren sich in der Erkennung des unstrukturierten Nachrichtentextes, herstellerspezifischen Informationen und der Ausgabe heraus. Durch manuelle Interaktionen des Anwenders können die Resultate jedoch optimiert werden. Die Software erfüllt somit die erwarteten Funktionen, ist aber für einen effizienten Einsatz in der Praxis noch nicht ausgereift.



Diese Grafik beschreibt den Ablauf der erstellten Anwendung. Die Log-Patterns werden anhand der eingelesenen Konfiguration identifiziert. Der Parser-Regex setzt sich zusammen aus den deduzierten Patterns und der Logdatei.