ABSTRACT

With the development of modern software systems, their monitoring, administration and maintenance are becoming increasingly difficult. Current approaches to system administration are based primarily on rules and guidelines that are shaped by experts in this area. Due to the increasing complexity and constantly changing environment, the process requires more and more time and is prone to errors. Accordingly, there is a need for an automatic and efficient system maintaining and monitoring solution based on the analysis of log data.

As part of this work, a software solution for unsupervised anomaly detection is developed based on textual log data from IBM System Automation for z/OS. The log data has information about the activities of software resources. An essential part of this work is the feature engineering, where the necessary characteristics of the data are worked out. Feature engineering for anomaly detection poses a challenge because the anomaly detection problem is unsupervised. The feature engineering results in two types of representation of the features, which are then examined for anomalies using existing data mining methods. Three different anomaly detection models are used. The first model is based on distance calculation, the second on Bayesian statistics and the third on an artificial neural network (LSTM). Using the methods, anomaly indicators will be shown and highlighted in the log.

Finally, the anomaly detection models and their results about the possibility to detect anomalies and interpretability will be discussed.

Keywords: Anomaly detection, Log analysis, Feature Engineering, Distance, Bayesian statistics, Long Short-Term Memory

ZUSAMMENFASSUNG

Mit der Entwicklung moderner Softwaresysteme werden ihre Überwachung, Verwaltung und Pflege zunehmend schwieriger. Aktuelle Ansätze zur Systemverwaltung beruhen vorwiegend auf Regeln und Richtlinien, die von Experten in diesem Bereich geprägt werden. Aufgrund der steigenden Komplexität und stets änderndes Umfelds verlangt der Prozess immer mehr Zeit und ist fehleranfällig. Dementsprechend besteht ein Bedarf an automatischer und effizienter Systemverwaltung, die auf der Analyse von Log-Daten basiert.

In Rahmen dieser Arbeit wird eine Softwarelösung für unüberwachte Anomalieerkennung anhand von textuellen Log-Daten der IBM System Automation für z/OS entwickelt. Die Log-Daten beinhalten die Informationen über Aktivitäten von Softwareressourcen. Ein wesentlicher Teil dieser Arbeit ist die Merkmalsextraktion, bei der die notwendigen Charakteristiken der Daten herausgearbeitet werden. Die Merkmalsextraktion für die Anomalieerkennung stellt eine Herausforderung dar, da das Anomalieerkennungsproblem nicht überwacht ist. Die Merkmalsextraktion resultiert in zwei Darstellungstype der Merkmale, die danach mittels existierenden Methoden des Data-Minings auf Anomalien untersucht werden. Es werden drei verschiedene Anomalieerkennungsmodelle eingesetzt. Das erste Modell basiert auf der Distanzberechnung, das zweite auf Bayes'scher Statistik und das dritte auf einem künstlichen neuronalen Netz (LSTM). Anhand der eingesetzten Methoden werden Anomalieindikatoren aufgezeigt und im Log hervorgehoben.

Abschließend werden die eingesetzten Anomalieerkennungsmodelle und ihre Ergebnisse bezüglich der Anomalieerkennung und der Interpretierbarkeit diskutiert.

Schlagwörter: Anomalieerkennung, Loganalyse, Merkmalsextraktion, Distanz, Bayes'sche Statistik, Long Short-Term Memory