Nowadays, big data systems (e.g., Hadoop and Spark) are being widely adopted by many domains for offering effective data solutions, such as manufacturing, healthcare, education, and media. A common problem about big data systems is an anomaly, i.e., a status deviated from normal execution, which decreases the performance of computation or kills running programs. It is becoming a necessity to detect anomalies and analyze their causes. An effective and economical approach is to analyze system logs. Big data systems produce numerous unstructured logs that contain buried valuable information. However manually detecting anomalies from system logs is a tedious and daunting task.

This dissertation proposes four approaches that can accurately and automatically analyze anomalies from big data system logs without extra monitoring overhead. Moreover, to detect abnormal tasks in Spark logs and analyze root causes, we design a utility to conduct fault injection and collect logs from multiple compute nodes. (1) Our first method is a statistical-based approach that can locate those abnormal tasks and calculate the weights of factors for analyzing the root causes. In the experiment, four potential root causes are considered, i.e., CPU, memory, network, and disk I/O. The experimental results show that the proposed approach is accurate in detecting abnormal tasks as well as finding the root causes. (2) To give a more reasonable probability result and avoid ad-hoc factor weights calculating, we propose a neural network approach to analyze root causes of abnormal tasks. We leverage General Regression Neural Network (GRNN) to identify root causes for abnormal tasks. The likelihood of reported root causes is presented to users according to the weighted factors by GRNN. (3) To further improve anomaly detection by avoiding feature extraction, we propose a novel approach by leveraging Convolutional Neural Networks (CNN). Our proposed model can automatically learn event relationships in system logs and detect anomaly with high accuracy. Our deep neural network consists of logkey2vec embeddings, three 1D convolutional layers, a dropout layer, and max pooling. According to our experiment, our CNN-based approach has better accuracy compared to other approaches using Long Short-Term Memory (LSTM) and Multilayer Perceptron (MLP)on detecting anomaly in Hadoop Distributed File System (HDFS) logs. (4) To analyze system logs more accurately, we extend our CNN-based approach with two attention schemes to detect anomalies in system logs. The proposed two attention schemes focus on different features from CNN's output. We evaluate our approaches with several benchmarks, the attention-based CNN model shows the best performance among all state-of-the-art methods.

Major: Computer Science

Educational Career:
Bachelor's of Software Engineering, BS, 2012, Taiyuan University of Technology
Master's of Software Engineering, MS, 2015, Tianjin University
Master's of Computer Science, MS, 2015, University of Wyoming

Committee in Charge:
Liqiang Wang, Chair, Department of Computer Science, UCF
Wei Zhang, Department of Computer Science, UCF
Shaojie Zhang, Department of Computer Science, UCF
Dazhong Wu, Department of Mechanical and Aerospace Engineering, UCF

Approved for distribution by Liqiang Wang, Committee Chair, on June 15, 2019.

The public is welcome to attend.