A Generator of Synthetic Access Logs that Contain Realistic User Behavior Patterns

Duane Bender, Weina Ma, Kamran Sartipi
Department of Electrical, Computer and Software Engineering
University of Ontario Institute of Technology
Oshawa, ON, L1H 7K4, Canada
{Duane.Bender, Weina.Ma, Kamran.Sartipi}@uoit.ca

Abstract

Generating high quality synthetic data for testing algorithms and system implementations is challenging. This research designed and developed a tool called QUAlity Synthetic Information Log Generator (Quasi-Log) to facilitate the development and testing of a new series of Information Discovery Systems (IDS) that focus on detecting User Behavior Patterns to improve the quality, security and potential profitability of systems. Quasi-Log is a powerful and highly configurable synthetic event log generation engine with advanced support for semantic attributes and value sets, attribute correlations, event sequence pattern definitions and user behaviour pattern definitions. In this paper we discuss the design of Quasi-Log and provide empirical results to demonstrate that our tool is effective at generating semi-random log files that exhibit these interesting properties.

Keywords: testing, quality assurance, dynamic analysis, pattern mining, log mining, discrete event simulation, synthetic data generation

I. Introduction

Analyzing log files reveals how users are interacting with the features of a system. This user behavior analysis provides valuable information to system operators which can be used to make system improvements, ultimately leading to improved security and resource access policies, a more satisfied user community and improved financial performance. Analyzing a log-file can provide indications of the quality of the software through feature analysis, as the logs will reveal what features of the software are being used. Unfortunately there is a general lack of access to real-world log files in critical domains such as healthcare, the military and other security-sensitive industries, thus creating synthetic datasets is the only pragmatic approach to test advances in user behavior analysis in those areas. Our tool has been designed in a general purpose way, allowing for pluggable modules which can exhibit domain-specific behavior as desired by the system analyst.

Modern organizations continue to become increasingly more interconnected. As the transactional volume in these systems increases, monitoring complex system activity such as user behaviour manually is no longer practical and requires automated techniques. Data mining systems such as association mining and sequence pattern mining systems are able to extract meaningful groupings and patterns from large volumes of transactional data. An active area of research includes new and refined algorithms to mine transactional data for user behavior and this tool supports such research.

Creating high quality synthetic data to simulate distributed systems is a challenge. We define "High quality" synthetic data as data that exhibits some realistic properties such as: event frequency that mimics normal work days and busy system times; attribute values that contain statistical biases towards certain types of operations (more reads than writes, for example); highly associated attribute values (specific users working in specific locations for example); and a mechanism for including common user behaviour patterns. Quasi-Log supports these features and more.

New data mining algorithms optimized to detect User Behavior require testing to determine the effectiveness of the approach. When synthetic data is used in the testing the effectiveness of the algorithms can be explicitly measured, i.e. User Behaviour Patterns that are inserted into the event log can be detected by the User Behavior Mining algorithm. The synthetic logs can also be used for benchmarking of production-level log databases. Ultimate benefits of this approach include improving the quality of the system by analyzing how the users use the system;