Simulation of Real-world Event Repositories for Evaluation of Data Analytics Solutions: Case of User Behavior Pattern Recovery

Hassan Sharghi, Weina Ma, Kamran Sartipi

***Abstract*—Due to the lack of access to the real-world event-log repositories in critical domains such as healthcare and banking, the evaluation and maintenance of data analytics algorithms has become a challenge. Generating synthetic log repositories that simulate a variety of complex real-world event-log repositories will be an effective way of producing benchmarks to evaluate data analytics algorithms using information retrieval metrics. As an important case study for such synthetic log repository, we populate an event-log repository with complex user-behavior instances in the healthcare domain, where the behavior is defined as a sequence of events by a user. Since user behavior has a complex nature, we defined a user-behavior pattern language (BPL) that allows the domain experts to represent both the desired behavior patterns that is used by the log generator engine, and for defining a target user-behavior pattern to be searched in the generated log repository. We use constraint-based approximate event-pattern matching techniques to search and identify the instances of the target pattern in the repository. In this paper, we introduce our BPL, present our event-log generator engine and the produced log repository, and use our pattern matching algorithm to identify the extracted user behaviours in the repository to show the practicality and usefulness of our proposed framework.**

***Index Terms*—User behavior; Log generator; Log repository; Event pattern; Pattern matching; Behavior language.**

I. Introduction

Large distributed information systems are increasingly vulnerable against insider-attacks and they are harder to maintain due to complexity of services, sensitive data resources, and service integration of incorporating enterprises with possible conflicting business rules. Ensuring the quality of services, proper resource utilization, and information integrity requires high-level system evaluations techniques based on monitoring and controlling the system usage. In such large distributed system, hundreds or thousands of concurrent users generate huge and complex dataset of service transactions which must be investigated all together to identify possible malicious user behaviours and overloaded resource utilizations. Most user activity monitoring systems (e.g., tracking suspicious credit card usage) track individual user's behavior and miss the overall dynamic usage patterns of users. On the other hand, different data analytics algorithms are developed based on event-log repositories which need a benchmark environment with known datasets to be tested and validated for correct operations and results, before they are applied on sensitive real-world production datasets. We have developed a comprehensive framework for two complementary purposes: i) generating a realistic and customizable synthetic event-log repository; and ii) extracting common and anomaly user behavior patterns from an event-log repository.

Behavior as a concept is broadly used in different context encompassing social, business, economic, and cultural domains. However, the lack of systematic and comprehensive methodologies and tools have prevented the advancements towards utilizing knowledge hidden in user behaviours [1]. Designing an efficient and scalable framework for monitoring and processing behavior patterns has been a major research interest in recent years to provide a mechanism for the enterprise to monitor the behavior patterns and anomalous behaviours of their customers. Behavior data are becoming a valuable asset to be carefully analyzed in order to reveal its explicit and implicit knowledge that cannot be attained just through recorded transactional data. Moreover, appropriate presentation of behavior analysis results is valuable for administrators to enable them to make decisions properly. Behavior modeling and representation attempt to develop representation languages and tools based on formal methods and techniques. The language helps the analyst to illustrate attributes, primitive events, semantics, constraints, and behavior patterns in a detailed and precise manner. The correlations between events based on attribute values along with constraints constitute the semantic of behavior pattern.

We developed a dataset generator toolkit (*EventGenerator*) for controllable dataset generation, suitable for unbiased evaluation of user behavior pattern recovery algorithms. The behavior pattern representation defines a scenario as a behavior pattern based on sequencing, timing and association rules, which allows data analysts to design interesting features and patterns that will be injected into the dataset. The generator creates datasets that are controlled by data size, data distribution, and the designed behavior patterns. Without such a generator it is impossible to test, validate and calibrate the algorithms that explore a system production dataset, as the