Weina Ma
Kamran Sartipi
{Weina.Ma, Kamran.Sartipi}@uoit.ca

Synthesizing Scenario-based Dataset for User Behavior Pattern Mining

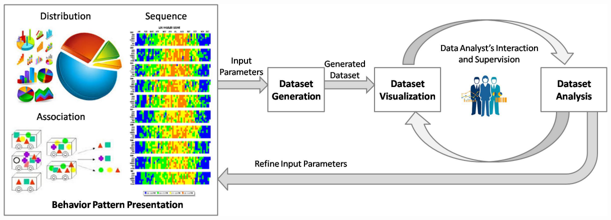
Motivations

- ➢ Identifying user behavior patterns from audit logs is valuable for system security of monitoring authorized users.
- ➢ Due to restricted access to production event-logs, security and privacy issues, and high costs of real datasets, synthetic event-log datasets are crucial in designing and evaluating data analytics approaches.
- ➢ A controlled event-log simulation environment provides the data analysts various synthetic dataset containing embedded interesting patterns and features. The produced testing datasets reduce the algorithm evaluation time.

Behavior Pattern

```
{
        {
            "id": "P-00001",
            "description": "typical doctor's workflow in radiology
            department",
            "context": {
               "role": "doctor",
               "department": "radiology"
            }
            "sequence": {
               "action":  "create an order",
               "action":  "read historical exams",
               "action":  "create an exam",
               "action": "create a report"
            },
            "interval": "within 3 days",
            "support": "10%"
        }
        {
            "id": "P-00002",
            "description": "daily nursing ward-round",
            "context": {
               "role": "nurse"
            }
            "sequence": {
               "location":  "ward-A",
               "location":  "ward-B",
               "location":  "ward-C"
            },
            "interval": "daily",
            "support": "15%"
        }
}
```

Architecture



We developed an interactive data exploration environment to such a design-generate-visualize-analyze-optimize process.

- ➢ ***Design***: statistical characteristics (distribution), association pattern, sequence pattern
- ➢ ***Generate***: produce a dataset that contains predefined attributes and patterns
- ➢ ***Visualize***: extract simplified workable information from generated dataset
- ➢ ***Analyze***: verify the differences between generated dataset and expected dataset
- ➢ ***Optimize***: refine input parameters

Proposed Approach

We proposed a synthetic event-log generator that effectively assists data analysts in designing scenario-driven event-logs with embedded user behavior patterns, and visually analyzing the quality of the generated datasets. The toolkit includes three layers:

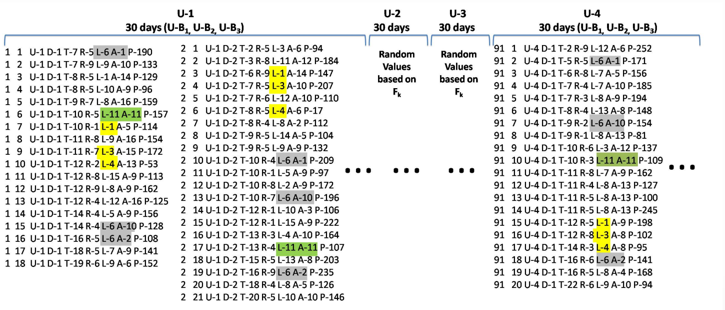
- ➢ ***Behavior pattern representation layer***: allows data analysts to design interesting features and patterns that will be injected into the dataset.
- ➢ ***Dataset generation layer***: creates datasets that are controlled by data size, data distribution, and the designed behavior patterns.
- ➢ ***Dataset visualization layer and analysis layer***: provides an interactive exploration environment for visual analysis of the quality of generated datasets.

Generator Algorithm

```
Algorithm: dataset-generator
Input: $A$, $V$, $B$, $avg$. $D_{All}$, $U_{All}$
Output: $E$

I = Ø, S= Ø , E = Ø , U = Ø
for $B_j$ in $B$ do begin
    association pattern $i_j$ = $B_j$-$c$
    sequence pattern $s_j$=$B_j$-$s$
    apply time constraint $B_j$-$t$ to $s_j$
    randomly select $B_j$-$sup$ users $U$-$B_j$
end for
for $a_k$ in $A$ do begin
    build biased random value select function $F_k$
end for
for each user $u_o$ in $U_{All}$ do begin
    for each day $d_p$ in $D_{All}$ do begin
        $x$ = randomly selected an integer around $avg$
        generate an empty event sequence $u_o$-$d_p$-$s_x$
        for $U$-$B_j$ in $U$ do begin
            if $u_o$ exists in $U$-$B_j$ then
                insert constraint-based sequence pattern $s_j$ to
                $u_o$- $d_p$-$s_x$
                insert association pattern $i_j$ to $u_o$-$d_p$-$s_x$
            end if
        end for
        for each event $e_l$ in $u_o$-$d_p$-$s_x$ do begin
            for each empty attribute $a_k$ in $e_l$ do begin
                call function $F_k$ to assign random value $v_{lk}$ to $e_l$
            end for
        end for
    end for
end for
```

Dataset Generator Output



U-1
30 days (U-B_1, U-B_2, U-B_3)

```
1  1  U-1 D-1 T-7 R-5 L-6 A-1 P-190
1  2  U-1 D-1 T-7 R-9 L-9 A-10 P-133
1  3  U-1 D-1 T-8 R-5 L-1 A-14 P-129
1  4  U-1 D-1 T-8 R-5 L-10 A-9 P-96
1  5  U-1 D-1 T-9 R-7 L-8 A-16 P-159
1  6  U-1 D-1 T-10 R-5 L-11 A-11 P-157
1  7  U-1 D-1 T-10 R-1 L-1 A-5 P-114
1  8  U-1 D-1 T-11 R-8 L-9 A-16 P-154
1  9  U-1 D-1 T-11 R-7 L-3 A-15 P-172
1 10  U-1 D-1 T-12 R-2 L-4 A-13 P-53
1 11  U-1 D-1 T-12 R-8 L-15 A-9 P-113
1 12  U-1 D-1 T-12 R-9 L-8 A-9 P-162
1 13  U-1 D-1 T-12 R-4 L-12 A-16 P-125
1 14  U-1 D-1 T-14 R-4 L-5 A-9 P-156
1 15  U-1 D-1 T-14 R-4 L-6 A-10 P-128
1 16  U-1 D-1 T-16 R-5 L-6 A-2 P-108
1 17  U-1 D-1 T-18 R-5 L-7 A-9 P-141
1 18  U-1 D-1 T-19 R-6 L-9 A-6 P-152
```

```
2  1  U-1 D-2 T-2 R-5 L-3 A-6 P-94
2  2  U-1 D-2 T-3 R-8 L-11 A-12 P-184
2  3  U-1 D-2 T-6 R-9 L-1 A-14 P-147
2  4  U-1 D-2 T-7 R-5 L-3 A-10 P-207
2  5  U-1 D-2 T-7 R-6 L-12 A-10 P-110
2  6  U-1 D-2 T-8 R-5 L-4 A-6 P-17
2  7  U-1 D-2 T-8 R-4 L-8 A-2 P-112
2  8  U-1 D-2 T-9 R-5 L-14 A-5 P-104
2  9  U-1 D-2 T-9 R-5 L-9 A-9 P-132
2 10  U-1 D-2 T-10 R-4 L-6 A-1 P-209
2 11  U-1 D-2 T-10 R-1 L-5 A-9 P-97
2 12  U-1 D-2 T-10 R-8 L-2 A-9 P-172
2 13  U-1 D-2 T-10 R-7 L-6 A-10 P-196
2 14  U-1 D-2 T-12 R-1 L-10 A-3 P-106
2 15  U-1 D-2 T-12 R-1 L-15 A-9 P-222
2 16  U-1 D-2 T-13 R-3 L-4 A-10 P-164
2 17  U-1 D-2 T-13 R-4 L-11 A-11 P-107
2 18  U-1 D-2 T-15 R-5 L-13 A-8 P-203
2 19  U-1 D-2 T-16 R-9 L-6 A-2 P-235
2 20  U-1 D-2 T-18 R-4 L-8 A-5 P-126
2 21  U-1 D-2 T-20 R-5 L-10 A-10 P-146
```

• • •

U-2
30 days

Random Values based on F_k

• • •

U-3
30 days

Random Values based on F_k

• • •

U-4
30 days (U-B_1, U-B_2, U-B_3)

```
91  1  U-4 D-1 T-2 R-9 L-12 A-6 P-252
91  2  U-4 D-1 T-5 R-5 L-6 A-1 P-171
91  3  U-4 D-1 T-6 R-8 L-7 A-5 P-156
91  4  U-4 D-1 T-7 R-4 L-7 A-10 P-185
91  5  U-4 D-1 T-7 R-3 L-8 A-9 P-194
91  6  U-4 D-1 T-8 R-4 L-13 A-8 P-148
91  7  U-4 D-1 T-9 R-2 L-6 A-10 P-154
91  8  U-4 D-1 T-9 R-1 L-8 A-13 P-81
91  9  U-4 D-1 T-10 R-6 L-3 A-12 P-137
91 10  U-4 D-1 T-10 R-3 L-11 A-11 P-109
91 11  U-4 D-1 T-11 R-8 L-7 A-9 P-162
91 12  U-4 D-1 T-11 R-4 L-8 A-13 P-127
91 13  U-4 D-1 T-11 R-5 L-8 A-13 P-100
91 14  U-4 D-1 T-11 R-5 L-8 A-13 P-245
91 15  U-4 D-1 T-12 R-5 L-1 A-9 P-198
91 16  U-4 D-1 T-12 R-8 L-3 A-8 P-102
91 17  U-4 D-1 T-14 R-3 L-4 A-8 P-95
91 18  U-4 D-1 T-16 R-6 L-6 A-2 P-141
91 19  U-4 D-1 T-16 R-5 L-8 A-4 P-168
91 20  U-4 D-1 T-22 R-6 L-9 A-10 P-94
```

• • •

A slice of generated event dataset for 4 users where the average events per user per day is 20. Users U-1 and U-4 are selected for insertion of 3 behavior patterns B1, B2 and B3, which are highlighted with different colors.

Implementation

Dataset Design

- ➢ Design an event dataset to simulate user-system interactions in distributed medical imaging systems.
- ➢ Each event has 6 attributes, where *Event=<User, Location, Action, Patient, Date, Time>.*
- ➢ Table II defined attribute distributions.
- ➢ Table III defined 9 typical user behavior patterns that constitute ordering, timing, and sequencing.
- ➢ Produced 30,000 events with randomly selected attribute values but following predefined distribution; predefined behavior patterns are inserted into the events.

TABLE II. ATTRIBUTE DISTRIBUTION DEFINITION

Attribute	Rep	Domain	Type	Mu	Sigma
User	U-	100	Random	-	-
Location	L-	15	Normal	8	4
Action	A-	16	Normal	8	3
Patient	P-	300	Normal	150	50
Date	D-	30	Normal	15	5
Time	T-	24	Normal	11	4

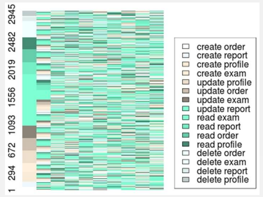
TABLE III. ANALYST DEFINED BEHAVIOR PATTERNS

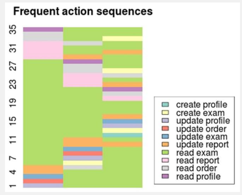
Pattern Id	Sequence	Support
P-00001	office-1-Juravinski-Hamilton, office-3-Juravinski-Hamilton, office-4-Juravinski-Hamilton	30%
P-00002	office-3-Lakeridge-Oshawa, office-4-Lakeridge-Oshawa, office-5-Lakeridge-Oshawa	25%
P-00003	office-2-McMaster-Hamilton, office-1-McMaster-Hamilton, office-3-McMaster-Hamilton	20%
P-00004	read exam, read report, update report	30%
P-00005	read exam, read order, create exam	25%
P-00006	create profile, read profile, update profile	20%
P-00007	11:00, 12:00, 14:00	30%
P-00008	10:00, 11:00, 12:00	25%
P-00009	14:00, 15:00, 16:00	20%

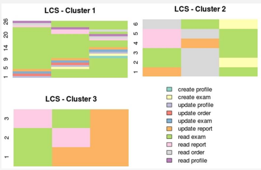
Visually Analysis of Generated Dataset

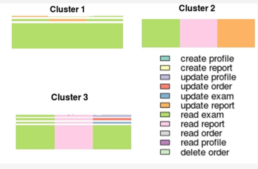
We developed a toolkit that can produce the following visual graphs for analyzing the dataset:

- ➢ ***Sequence overview***: sequence is an ordered list of events performed by one person per day.
- ➢ ***Frequent sequential patterns***: are subsequences that appear frequently among all user sequences
- ➢ ***Clustering based on sequence similarity***: divides the frequent sequential patterns into a number of clusters
- ➢ ***Clustering representatives***: explores the representative patterns of each cluster


create order
create report
create profile
create exam
update profile
update order
update exam
update report
read exam
read report
read order
read profile
delete order
delete exam
delete report
delete profile


Frequent action sequences
create profile
create exam
update profile
update order
update exam
update report
read exam
read report
read order
read profile


LCS - Cluster 1
LCS - Cluster 2
LCS - Cluster 3
create profile
create exam
update profile
update order
update exam
update report
read exam
read report
read order
read profile


Cluster 1
Cluster 2
Cluster 3
create profile
create report
update profile
update order
update exam
update report
read exam
read report
read order
read profile
delete order