BP-Meet-IoT 2022 Challenge

Francesco Leotta, Massimo Mecella
Estefanía Serral, Victoria Torres

April 6, 2022

1 Introduction

The 2022 edition of the BP-Meet-Iot Challenge requires participants to analyze smart home datasets performing tasks including but not limited to:

- IoT-aware business process modeling;
- Log preprocessing for process mining;
- Conformance analysis of logs with models;
- Action/ADL/Habit prediction;
- Dataset visualization;
- Resource identification;
- ...

the challenge proposes in particular two different kinds of datasets:

- Real-World Smart Home Datasets commonly used in the smart home and ambient intelligence community listed in Section 2;
- A simulated dataset, see Section 3 obtained by executing documented human processes in a virtual environment.

2 Real-World Smart Home Datasets

For this year challenge we propose you to use the wide collection of datasets provided by the CASAS project[1].

As an example, the Aruba dataset consists of a sensor log containing raw sensor measurements collected in a smart home inhabited by an adult woman for 220 days. The floorplan and the set of available sensors of this installation are shown in Figure 1. In particular, the environment contains:

[1] `http://casas.wsu.edu/datasets/`

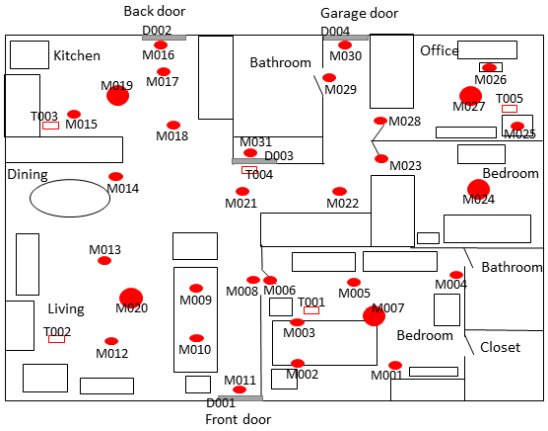

Figure 1: Aruba installation from CASAS project.

- *Presence InfraRed (PIR) Sensors* represented by small and large red ellipses and labeled as MXXX;
- *temperature sensors* represented by red empty rectangles and denoted with a label of the form TXXX;
- *door switch sensors* represented by gray rectangles and denoted with a label of the form DXXX.

Summarizing, this smart home is equipped with 31 PIR motion sensors used to detect the presence of a person moving under them; 5 temperature sensors to measure the temperature of a specific room; 3 door switch sensors to detect when the resident is entering/leaving home.

PIR sensors are attached to the ceiling of the house facing downwards, so they trigger as soon as a person enters inside of their detection area. For example, sensors M002 and M003 trigger whenever a person sits on the corresponding side of the bed, whereas sensors M009 and M010 provide the same information with respect to the sofa. This smart environment contains two different kinds of PIR sensors; large area PIRs, represented by large ellipses, featuring large detection cones, are used to generically detect the area of the house the person is occupying at a specific moment; small area PIRs, represented by small ellipses, potentially indicate the inhabitant interacting with a device in that specific and restricted portion of the house.

The dataset is partially labeled with the start and the end markers of activities performed by the resident, meaning that for a subset of measurements we know the activity correspondent to those activations of sensors. The activities available in the dataset, with the correspondent number of their occurrences, are the following ones: *meal preparation* (1606), *relax* (2910), *eating* (257), *work* (171), *sleeping* (401), *wash dishes* (65), *bed to toilet* (157), *enter home* (431), *leave home* (431), *housekeeping* (33) and *resperate* (6).

Other datasets from the CASAS project can be used. They vary in terms of:

- **Inhabitants**. The number of inhabitants involved in the dataset acquisition process. An important aspect is whether the users are present in the smart space at the same time or in separate sessions.
- **Activities**. The number of different activities that the dataset captures.
- **Labeling**. In most datasets, participants are required to perform, or to label, a certain set of activities. Having labeled activities allows to create specific models to perform activity-specific tasks such as recognition and prediction. Activities can be performed by single users or collaboratively by multiple users. Labels usually refer to what activity/ies is/are performed in a specific time range and are applied *(i)* by researchers in case of controlled experiments (i.e., experiments where participants are asked to perform specific tasks), or *(ii)* by participants through specific labeling tools. Finally, the fraction of the dataset which is actually labeled may vary.
- **Length**. The size of a dataset in terms of days of acquisition.

3 Business Process Based Dataset

The challenge also provides a set of simulated datasets that describe habits, activities and actions performed by humans inside a physical environment, together with the sensor measurements detected. In particular:

- Actions are atomic interactions with the environment or a part of it (e.g., a device). With respect to the common BPM terminology, they correspond to tasks of a process model, whose execution is witnessed by events in an event log.
- Activities are combinations of actions. With respect to the common BPM terminology, they correspond to processes, whose execution is witnessed by traces in an event log.
- Habits are sequence or interleaving of activities that happen in specific contextual conditions (e.g., what the user does each morning between 08:00 am and 10:00 am). A habit is similar to an activity, but here the focus is on the routine rather than on the final goal. With respect to the common BPM terminology, they correspond to macro-processes made up by several processes.

The provided datasets are organized according to two different scenarios:

- A scenario, denoted as `d21p1`, describing the daily habits (which happen from 0:00am to 11:59pm) of a single user for 21 consecutive days;

- A scenario, denoted as `d21p2`, describing the daily habits (which happen from 0:00am to 11:59pm) of two users for 21 consecutive days.

All the scenarios cover three consecutive weeks, starting on a Monday (the fictitious date 2020-03-16) and concluding on a Sunday. As it will be explained later, this is important as weekend habits are different than working days ones.

Three different datasets are provided for each scenario representing human life at different levels of granularity:

- A dataset (stored in the EventLogXES.xes file) providing information about the activities performed by humans. For each of these activities the sequence of actions is provided;
- A dataset (stored in the EventLogXESNoSegment.xes file) only providing actions performed in the environment. This dataset is equivalent to the previous one, but here actions are not explicitly segmented into activities; noteworthy, in order to connect this dataset to the previous one, action (event) ids are equivalent;
- A dataset (stored in the SensorXES.xes file) containing sensor measurements. Here, sensor measurements are represented as events of a single trace.

3.1 Description of the environment

The physical environment where the users are is depicted in Figure 2.

This map is the faithful reproduction of the Aruba installation from the CASAS project [2]. Blue lines represent walls, green icons represent position sensors with their respective orientation, black icons represent landmark positions including wall limits, denoted as pxx (with xx going from 1 to 43), and devices and position of interest, including:

- Indoor positions: bathtub, exercise place, wardrobe, bed, windows, tv chair, dining table, dining chair, computer, computer chair, chair, kitchen sink, shoe shelf, fridge, micro, kitchen shelf and oven;
- Outdoor positions: Start, entrance, outside and workplace. Entrance is the point where the door house is. Outside and workplace are generic positions introduced for certain activities. Start is instead a fake position used when no information about a human position are available.

Positions of places and position sensors on the maps are expressed in terms of x (horizontal) and y (vertical) coordinates. A description file, `house.txt` is provided together with the datasets, containing the exact coordinates of places, walls and position sensors depicted in Figure 2.

Measurements from the following may be considered/captured/included in the datasets:

[2] cf. `http://casas.wsu.edu/datasets/`

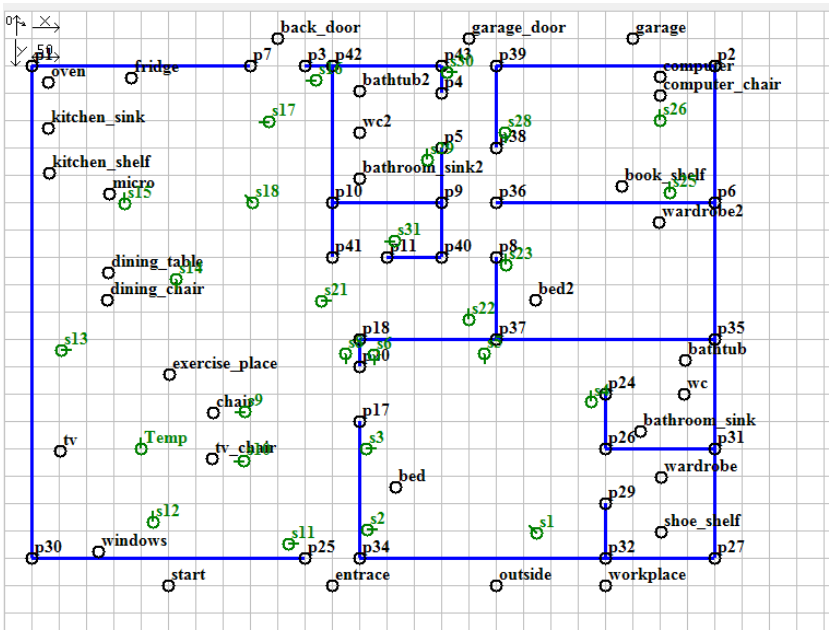

Figure 2: The house map

- For each user, a position (denoted as `position`) is provided. This position integrates the measurements provided by position sensors in the environments in order to provide, for each user the x and y coordinates occupied inside the environment. Values provided are strings in the form ẍ yː For example 300 1050 must be interpreted as $x = 300$ and $y = 1050$. Both x and y are integers;

- A temperature sensor (denoted as `Temp`) obtained by readings temperature values from an external service, i.e., temperature values are not simulated but read from an external oracle. Values are float numbers representing Celsius degrees;

- A home airing level sensor (denoted as `Home_Aired`), providing the level of air in the environment. This sensor is simulated and influenced by the actions performed by the user (e.g., opening the window increases the airing level). This is a discrete sensor with integer (also negative) values;

- A bed pressure sensor (denoted as `bed_pressure`), providing information about how many persons are currently on the bed. This sensor value is increased by one every time somebody is on the bed. Values from this sensor are positive or zero integers;

- A window and a blind switch sensor (denoted respectively as `windows` and `blinds`), providing information about the state of the window and

corresponding blind. A value of 1 means the window or the blind is open;

- A sensor (denoted as `unwashed_dishes`) providing the number of unwashed dishes in the house. This is a positive or zero integer;
- A sensor (denoted as `food`) providing the level of food availability in the house. Values of this sensors are positive or zero integer values;
- A sensor (denoted as `air_Conditioning`) reporting the state of the air conditioning system. This sensor has value 1 if the air conditioning system is on, and 0 otherwise;
- A sensor (denoted as `Home_Presence` reporting the number of persons currently present in the house. The value is an integer greater or equal than zero;
- A switch sensor (denoted as `fridge_door_contact`) connected to the fridge. This sensor is 0 when the fridge is open, 1 otherwise;
- A sensor (denoted as `cooked_food`) reporting how much cooked food is available. Similarly to the `food` sensor these are positive or zero integer values;
- A sensor (denoted as `power_use`) reporting the current usage of power in the house. Power is represented with a non-physical positive or zero integer;
- A sensor (denoted as `water_use` indicating the current employment of water in the house. Water usage is represented with a non-physical positive or zero integer.

Sensor values are captured when a value change is detected, and are provided in the SensorXES.xes dataset.

3.2 Description of activities

Humans in the simulated environment perform activities based on need or on specific environmental conditions. The execution of each activity implies the execution of actions. A priority mechanism is applied in order to let certain urgent activities to pause the execution of a specific activity. Each activity is performed by a single human. Names associated to activities are self-explanatory and will not be discussed further in this guide. Some of the activities, are the result of the execution of the models depicted in Figure 3 where the formalism employed is the one introduced in cite [2, 1]. Here the root node denotes the name of the activity as reported in the event logs.

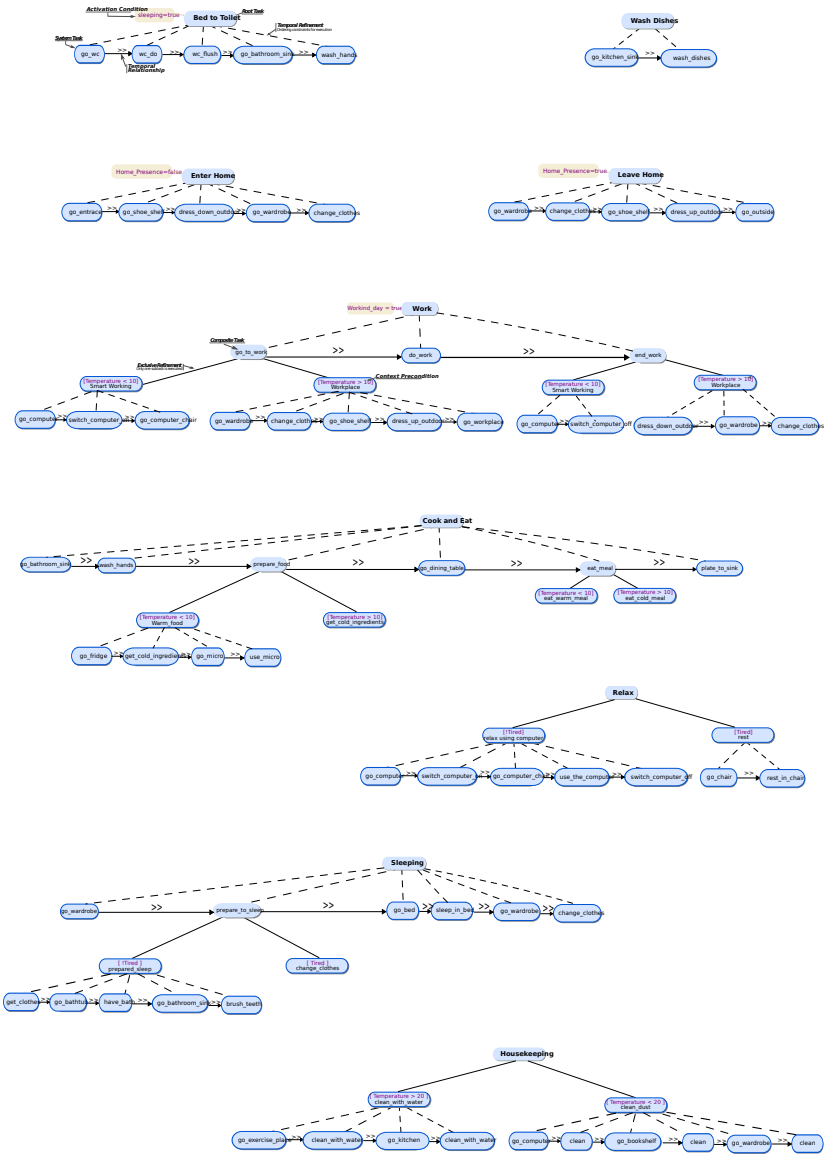

Figure 3: Task models

3.3 Clarifications on the XES format used in the datasets

Datasets are provided following the XES standard, but some clarifications are needed about specific aspects:

- In an XES log, usually instances of the same process are provided. Here, each daily (or morning) habit is the equivalent of process, where traces represent repetitions of specific activities. For this reason, for each activity trace, a name is provided;
- A trace ID is associated to each trace and an event ID is associated to each event. The event ID in the EventLogXES.xes and the EventLogXESNoSegment.xes of the same scenario are equivalent, so that the two datasets can be compared;
- Each event is associated with a resource that can be a human, if s/he has executed the action that causes the event, or the system, if the action has automatically been performed by the system;
- When possible, for each action, the start and complete events of the lifecycle are reported, so that the duration component of an action can be taken into account. When a start event is not available, only the complete event is reported, as such simulating an actual event log;
- Some events, denoted as noise, are functional to the simulation and can be ignored.

References

[1] Estefanía Serral, Pedro Valderas, and Vicente Pelechano. Addressing the evolution of automated user behaviour patterns by runtime model interpretation. *Software & Systems Modeling*, 14(4):1387–1420, 2015.

[2] Estefanía Serral Asensio, Pedro José Valderas Aranda, and Vicente Pelechano Ferragud. Supporting ambient assisting living by using executable context-adaptive task models. *International Journal On Advances in Software*, 7(1&2):77–87, 2014.