A Framework To Extract Personalized Behavioural Patterns of User's IoT Devices Data

Pradeep K. Venkatesh
Department of Electrical and Computer Engineering
Queen's University
Kingston, Ontario, Canada
15pkv@queensu.ca

Daniel Alencar da Costa
Department of Electrical and Computer Engineering
Queen's University
Kingston, Ontario, Canada
daniel.alencar@queensu.ca

Ying Zou
Department of Electrical and Computer Engineering
Queen's University
Kingston, Ontario, Canada
ying.zou@queensu.ca

Joanna W. Ng
IBM Watson Internet of Things
IBM Canada Ltd.
Markham, Ontario, Canada
jwng@ca.ibm.com

ABSTRACT

The growing trend of devices participation in Internet of Things (IoTs) platforms have created billions of IoT devices in both consumer and industrial environments. IoT devices form the network of devices connected to each other by communication technologies in different environments to monitor, collect, exchange, and to take actions. Due to the growth of IoT devices, it is cheaper and easily available so users started using these devices to achieve their personal goals, such as to reduce electricity cost at home. Existing research has proposed new interconnection implementation mechanisms for IoT devices to monitor environments by low cost systems. However, existing work does not investigate the historical data of IoT device usage to assist users in achieving their goals. In our research, we propose an engine that identifies the behavioural patterns of IoT device users. Our engine works in three steps: First, the engine uses a database to store the IoT devices usage data. Second, our engine prepares the data in a suitable model for data analysis. Finally, our engine analyses the represented data to extract user behavioural patterns. We perform an empirical study to evaluate our engine. Our results shows that users, on average, use less than 50% of their IoT devices at specific times and have a relatively small impact across other devices in the environment.

IBM CASCON 2017, Nov 2017, Markham, Ontario, Canada

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

KEYWORDS

Internet of Things (IoTs), Rule Extraction, Smart Home, Home Automation, Personalized Software Systems.

ACM Reference format:
Pradeep K. Venkatesh, Daniel Alencar da Costa, Ying Zou, and Joanna W. Ng. 2017. A Framework To Extract Personalized Behavioural Patterns of User's IoT Devices Data. In *Proceedings of IBM CANADA, Markham, Ontario, Canada, Nov 2017 (IBM CASCON 2017),* 9 pages.
https://doi.org/10.475/123_4

1 INTRODUCTION

The exponential growth of smart devices participation in the Internet of Things (IoTs) platform created more then 2 billion smart devices since 2006 and is expected to reach about 50 billion connected smart devices in 2020 (*e.g.*, [17, 22, 24]). The IoT devices are a network of internet-connected devices that collect and exchange data using embedded sensors. Due to the increasing growth of IoT devices, many companies have started investing to produce IoT devices for consumers, such as Amazon, Google, and Cisco. In addition, consumers (*i.e.*, users) use devices in their favour to achieve their personal goals, such as to save electricity cost at their homes.

Usually, IoT devices perform the following three tasks for users: 1) Sense and Monitor the environment (*e.g.*, monitor the room temperature); 2) Perform certain actions (*e.g.*, turn on/off lights); or 3) both 1 and 2 (*e.g.*, if room temperature reaches above 24°C then turn on the air conditioner). The IoT devices have their own user interface to let users take control of that particular device in a particular environment. Typically, users install multiple IoT devices in the desired environments, such as smart garage door, kitchen lights, and a foyer fan. However, the process of controlling each IoT device individually is very tiresome, since the manual effort from a user causes frustration (*e.g.*, [9]).

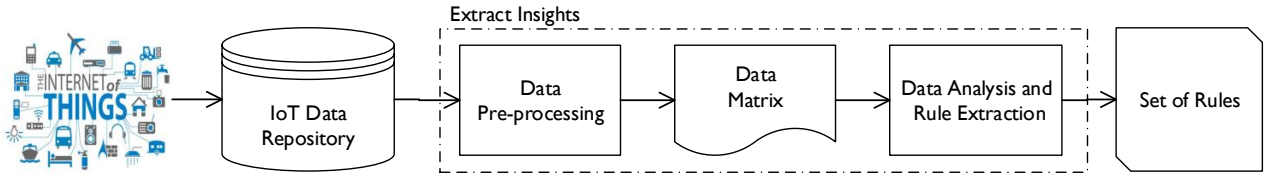

Figure 1: An overview of our behavioural extraction engine.

Research has been invested to help users achieve their personal goals using IoT devices (*e.g.*, [8, 13, 20, 27]). For example, existing research helps the identification of different interconnection mechanisms among the IoT devices to save energy in a smart home (*e.g.*, [20, 27]). Other line of research helps to better manage the resources of resource constraint IoT devices (*e.g.*, [8]). Kelly *et al.* [13] proposed a new implementation mechanism for IoT devices to monitor domestic conditions by means of low cost ubiquitous sensing systems. However, the existing prior work does not investigate the historical data of IoT devices usage to assist users in achieving their goals. For example, learning users' behavioural patterns from their IoT devices usage to automatically help users to achieve their personal goals.

In this paper, we propose a personalized behavioural extraction engine using the Apriori algorithm, a rule mining learning technique. Our engine infers behavioural patterns from IoT devices usage data. The extracted behavioural patterns can be used to help users to intelligently control their IoT devices with minimal user involvement. An example of a behavioural pattern is: *during evening hours, the kitchen lights are ON, while the garage doors are CLOSED.* Our approach learns these behavioural patterns and use them to control the IoT devices. Our engine works in three steps. First, the engine uses a database to store the IoT devices usage data. Second, our engine prepares the data in a suitable representational model for analysis. Finally, the engine analyses the represented data to extract user behavioural patterns. To evaluate our approach, we perform an empirical study using 4 users and 31 IoT devices usage data collected from a well-known IoT data publishing-subscription site (*i.e.*,dweet.io[1]). We derive the behavioural patterns for users'. In this study, we investigate the following research questions:

(RQ1.) ***What are the most used devices at a given time?*** To understand which IoT devices are most used by users at a given time helps us to identify behavioural patterns of IoT device usage and becomes a central point for home automation. (*e.g.*, garage door always open between 8am to 9am on weekdays). In total, we mine 35 behavioural rules of all the devices from 4 users. We observe that, on average, users use less than 50% of their IoT devices at specific times. Hence, users can concentrate on a limited number of devices when trying to control their environment.

(RQ2.) ***What is the relationship between the most used devices and the other devices in the environment?*** Studying the relationships between the most used devices and the other devices of the environment may be useful to give additional behavioural patterns (*e.g.*, one relationship shows that whenever the garage door is CLOSED the kitchen lights are ON). We observe, a small proportion of relationships among devices with a confidence interval between 50% – 80%. Therefore, users can use these identified relationships to intelligently take actions across other devices in the environment.

Paper organization. Section 2 summarizes the background of our work. Section 3 describes our proposed approach. Section 4 shows our empirical studies. Section 5 discusses the threats to validity of our work. Section 6 summarizes the related work. Finally, we conclude and provide directions for future work in Section 7.

2 BACKGROUND

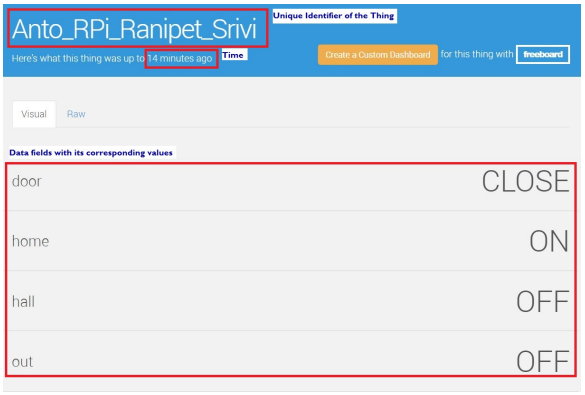

Figure 2: An annotated screen shot of the dweet.io, a published message of a device.

dweet.io is a simple publishing and subscribing site for users to store and share their device data in real-time. The device refers to machines, sensors, devices, robots, and gadgets.

[1]https://dweet.io/play/

The publishing data messages to the platform are referred as "dweets". The dweet.io site allows dweets to be up to 2,000 characters payload. The dweets can be easily accessed through a web based RESTful API.[2] Typically, users publish their device data for simple sharing, storage, and alerts purposes.

Figure 2 illustrates a dweet message posted on the dweet.io site. There are mainly four pieces of information: a unique identifier of the device, the data fields of the device and its corresponding values, and the time it was posted.

3 OUR PROPOSED ENGINE

In this section, we present our proposed behavioural extraction engine. We describe the components of our engine as well as which inputs and outputs are consumed and generated by them. We also discuss the necessary steps that our engine performs to extract behavioural patterns. Figure 1 provides an overview of how our engine works.

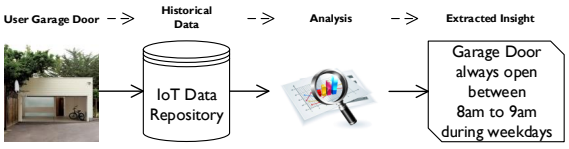

Figure 3: An example of how our engine works.

Our engine works in three simple steps. First, our engine collects the usage data for each IoT device. For example, the engine records the status of the device and its respective time-stamp (*e.g.*,{Device: Garage Door, Status: OPEN, Date: 01/01/2017, Time: 8am}). Next, in Step 2, our engine prepares the data in a suitable representational model for analysis (see *e.g.*, Equation 1). Finally, our engine analyses the representational data model to extract user behavioural patterns that are used to assist users in achieving their personal household goals. Figure 3 briefly illustrates the overall steps of how our behavioural extraction engine works.

3.1 IoT Data Collection

We developed a python crawler to access and download the web pages that have dweet posts from users. We parse the DOM tree[3] of a web page to extract the information of the dweet posts. The extracted dweet posts are cleaned and processed for data analysis with the following steps:

- Remove any code snippets other than raw data field containers from each dweet posts (*i.e.*, statements not enclosed with tag "<raw>...</raw>"). We filter out such an information because it is not necessary to extract an user's behaviour.
- Remove any data fields whose values are not in a form to convert to a binary form from each dweet posts, as these fields are not informative to our study. For example, temperature data field (*i.e.*, 21°C).
- Apply the conversion mechanism that maps data field values to their binary form. For instance, OPEN and CLOSE are converted to "1" and "0".

3.2 IoT Data Repository Design

Each IoT smart device is associated with a user, environment (*e.g.*, Home), and a set of states (*e.g.*,OPEN or CLOSE). To capture all these elements, we model the Entity-Relationship (ER) model [4] that is shown in Figure 4. The IoT data repository stores the usage of each IoT device in the format of transactions.

Each transaction (T) in the repository is represented in the form of:

$$T = \{DeviceId, UserId, User\_EnvironmentId, Status, Time\_Stamp\} \quad (2)$$

where DeviceId is a unique identifier for an IoT device. UserId is the name of the user who owns the IoT device while User_EnvironmentId is the environment where a device resides. Status is the current status of the device and Time_Stamp denotes the recorded date and time of the transaction.

3.3 Data Representation

To prepare the data for analysis, we present the data in the form of matrix using Equation 1, where each row denotes the time (*i.e.*, 9am, 10am, *etc.*) and each column represents a day. The vector in each cell represents the list of all devices in the environment with its corresponding status at the given day and time, such as Monday at 9am: {Garage Door: OPEN, Foyer Light: ON, Basement Light: ON}.

3.4 Behavioural Patterns Extraction

Our engine identifies the behavioural patterns of users' IoT device usage in a two-fold process. First, our engine measures each device usage by summarizing the frequency of possible device states at specific times in the environment. For example, Foyer lights are turned On 90% of times between 8am till noon. Second, our engine identifies the impact (*i.e.*, corresponding actions and relationships) among other IoT devices in the environment using Apriori algorithm. For example, the foyer lights are On between 8am till noon then the front door is Closed.

Apriori Algorithm. It is an association rule mining technique which mines if-then rules given a set of transactions [2]. For instance, given a set of transactions (Trans.) the rules

[2]https://dweet.io/get/latest/dweet/for/<my-thing-name>, where <my-thing-name> should be replaced with the assigned unique name of the device.
[3]https://www.w3.org/TR/DOM-Level-2-Core/introduction.html

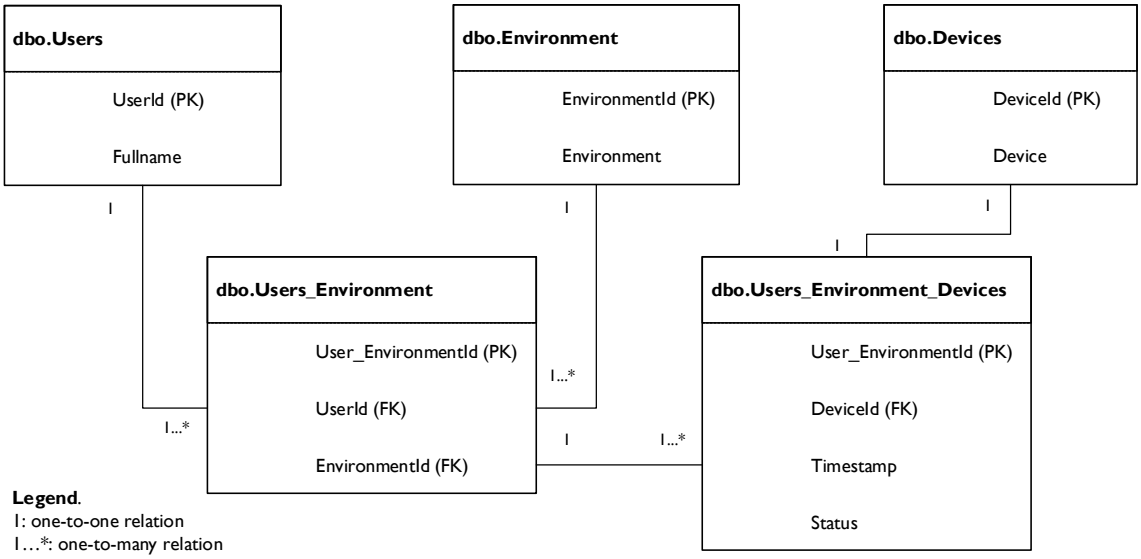

Figure 4: The entity-relationship diagram of IoT data repository.

Data Matrix Model =

$$\begin{bmatrix} & \mathbf{Day}_1 & \mathbf{Day}_2 & .. & \mathbf{Day}_n \\ \mathbf{T}_1 & (d1:ON, d2:ON, .., dn:OFF) & (d1:OFF, d2:ON, .., dn:OFF) & & (d1:ON, d2:ON, .., dn:OFF) \\ \mathbf{T}_2 & (d1:OFF, d2:OFF, .., dn:OFF) & (d1:ON, d2:ON, .., dn:OFF) & & (d1:OFF, d2:OFF, .., dn:OFF) \\ \mathbf{T}_3 & (d1:ON, d2:ON, .., dn:OFF) & (d1:ON, d2:ON, .., dn:OFF) & & (d1:ON, d2:OFF, .., dn:ON) \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{T}_n & (d1:OFF, d2:ON, .., dn:ON) & (d1:ON, d2:ON, .., dn:OFF) & & (d1:ON, d2:OFF, .., dn:OFF) \end{bmatrix} \quad (1)$$

are mined to form implication expressions in the form of if-then rules (*e.g.*, X1 $\implies$ Y1, where X1 and Y1 are disjoint itemsets on devices in our case). The formed if-then rules are extracted using the following two metrics: how often a rule is applicable in the given data and how frequently items in Y1 appears in transactions that contain X1 (*i.e.*, X1 and Y1 are IoT devices in the house-hold in our case).

Our engine's analysis steps are explained in depth in the upcoming research questions.

4 CASE STUDY

In this section, we present our case study setup and results. We describe the motivation, the approach and the findings of our two research questions.

4.1 Case Study Setup

IoT Data Repository. We create the IoT data repository with users' IoT devices usage data. We use a well-known publishing and subscribing site for IoT devices (*i.e.*,dweet.io) to collect the IoT devices real-time usage data for 4 users. The collected IoT devices of all users are from smart home environments. Table 1 shows the detailed summarizes of the IoT devices data collected for each user. The table describes the list of devices for each user with its corresponding description and possible device states (*e.g.*, Open or Close).

4.2 Research Question

***RQ1.* What are the most used devices at a given time?**

Motivation. Controlling the devices of an environment requires a great effort from users because controlling each device may be very tiresome for users. Therefore, it is important to identify which devices are used the most by users at a given time. This analysis allows us to learn their habits of devices usage. Users can use this knowledge to minimize their effort on controlling the devices in their environment. Further, it will become a central point for home automation

Analysis Approach. To identify which devices are used the most on specific times, we perform the following steps.

Table 1: List of studied IoT devices.

IoT Devices	States
(User 1): Ranipet House	
(D1). Front Door: provides the status of the front door.	(OPEN|CLS)
(D2). Outside Light: shows the current status of the outside light.	(ON|OFF)
(D3). Hall Light: shows the current status of the hall room light.	(ON|OFF)
(D4). Home Fan: shows the running status of the fan.	(ON|OFF)
(User 2): Alastair House	
(D1). Front Door: provides the status of the house's front door.	(OPEN|CLS)
(D2). Back Door: shows whether the back door is open or not.	(OPEN|CLS)
(D3). Garage Door: shows whether the garage door is open or not.	(OPEN|CLS)
(User 3): Azer House	
(D1). Family Room: provides the status of the family room light.	(ON|OFF)
(D2). Kitchen: indicates the status of the kitchen lights.	(ON|OFF)
(D3). Bedroom: indicates the status of the master bedroom light.	(ON|OFF)
(D4). Foyer: provides the status of the foyer room light.	(ON|OFF)
(User 4): Laabs House	
(D1). Bedroom L: shows the status of the bedroom light.	(ON | OFF)
(D2). Track Light Front: shows the status of the front track lights.	(ON | OFF)
(D3). Track Light Rear: shows the status of the back track lights.	(ON | OFF)
(D4). Downstairs L: shows the status of the downstairs hall light.	(ON | OFF)
(D5). Upstairs Light: shows the status of the upstairs hall light.	(ON | OFF)
(D6). Shower: shows the status of the master bathroom shower.	(ON | OFF)
(D7). Vanity L: the master bathroom vanity light status.	(ON | OFF)
(D8). Bathroom: shows whether the bathroom light is on or not.	(ON | OFF)
(D9). Washer: indicate whether the garage washer is active or not.	(ON | OFF)
(D10). Nursery Floor: the nursery lamp status.	(ON | OFF)
(D11). Porch: shows whether the porch front light is on or not.	(ON | OFF)
(D12). Entryway: indicates the status of the entry way hall light.	(ON | OFF)
(D13). Bedroom CL: shows the state of the bedroom corner lamp.	(ON | OFF)
(D14). Floor Lamp: indicates the hall floor lamp status.	(ON | OFF)
(D15). Kitchen: current status of the kitchen room light.	(ON | OFF)
(D16). Entry Table: hallway entry table lamp status.	(ON | OFF)
(D17). Downstairs Fan: downstairs bathroom fan status.	(ON | OFF)
(D18). Bathroom Fan: shows the master bathroom fan status.	(ON | OFF)
(D19). Bed LED: the master bedroom underbed LED strip status.	(ON | OFF)
(D20). Garage L: the garage fluorescent light status.	(ON | OFF)

Note:- CLS denotes the device state Close.

First, for each user, we arrange the IoT devices data in the IoT repository in a form of data matrix model as outlined in Section 3.3. In the data matrix model, each row represents the devices state in the environment at a specific time (*e.g.*, 9am) and each column represent the day of the week (*e.g.*, Sunday). Second, to measure the devices usage, we introduce a new metric DFreq (see Equation 3) to calculate the number of times the device remains in a specific state at given times.

$$DFreq = \frac{\sum_{Time=1}^{n}(\sum_{Day=1}^{m} DeviceStatus_{Day} : State)}{Total\ number\ of\ days} \quad (3)$$

*where the State represents the status of a device (*i.e.*, either On or Off, or Open or Close). In this research question, we aim to study the devices that are the most used (*i.e.*, device state either On or Open). Therefore, by default we set the State variable of the Equation 3 to be 1 (*i.e.*, device state either ON or Open).*

Third, to identify the devices that are used by users at specific times, we filter the data matrix by applying a 80/20 Pareto's principle [18] (*i.e.*, DFreq>=80% confidence interval threshold), where the 80% of data represents the remaining 20% of data population. In this way, we derive the most used devices by users on specific times.

Results. For each user, only 2 devices are being used the most in the environment at specific times. The most used devices are shown in Table 2. For instance, there are 2 out of 4 devices that are majorly used by User 1. Those devices are the Hall Light and Home Fan which are being used during Weekends from 8:45am to 11:45am and 9:45am to 10:45am, respectively.

In particular, we observe that for User 2 we cannot identify which devices that are mostly used, since those devices do not achieve DFreq>=80% in the studied period. Therefore, we eliminate those devices from further analysis.

The results presented in Table 2 shows the users' behavioural usage pattern for the most used devices only, we show the usage patterns for Weekdays and Weekends. However, if desired, the same technique could be applied across different time-frames to learn the behavioural patterns of users, such as each day of the week (*e.g.*, Monday), monthly (*e.g.*, April), and seasonally (*i.e.*, summer, fall, winter, and spring). The behavioural patterns can be used to help users by automatically taking actions on their most used devices. For example, our behavioural pattern for User1 shows that his/her Hall light is ON during Weekdays from 8:30 to 12:15pm. We can use this learned behavioural pattern to automatically turn ON the Hall light during this time-frame to assist User1.

On average, users' use less than 50% of their IoT devices at specific times in the environment.

Table 2: List of the most used IoT devices per users.

IoT devices	Time-frames	
	Weekdays	**Weekends**
(User 1)		
(D3). Hall Light	8:30 to 12:15pm	8:45am to 11:45am
(D4). Home Fan	8:30 to 11:30am	9:45am to 10:45am
(User 2)	-/-	
(User3)		
(D1). Family Room	1:15 to 1:30am	1 to 1:15am; 7:45 to 8am;
(D2). Kitchen	-/-	1:15pm to 1:45pm; 3:15pm to 4:30pm; 10 to 10:15pm
(D3). Bedroom	6:45pm to 7pm	3 to 3:15pm
(D4). Foyer	7:45pm to 8pm	3 to 3:15am; 12:15 t0 12:45pm
(User 4)		
(D1). Bedroom Light	10:45 to 11:30am	1:15am to 2:30am
(D2). Track Light Front	12pm to 12:15pm; 4pm to 4:15pm; 8:30pm to 3am	12am to 3am; 8:45pm to 12am
(D3). Track Light Rear	12pm to 12:30pm; 8:15 to 3am	12am to 3am; 8:45pm to 12am
(D11). Porch	10:15pm to 2:30am;	12 to 3am; 10:15pm to 12am
(D13). Bedroom CL	1:45 to 2am; 11am to 11:45am	12am to 2:30am; 10:30pm to 12am
(D15). Kitchen	12am to 11:45pm	12am to 11:45pm

RQ2. ***What is the relationship between the most used devices and the other devices in the environment?***

Motivation. In a given environment, users tend to use multiple IoT devices to achieve their personal goals. From RQ1, we observe the IoT devices that are mostly used by each user at specific times. However, these most used devices may impact the usage of other devices. For example, whenever the garage door closes the kitchen lights are turned ON. Therefore, in RQ2 we aim to study the impact (*i.e.*, corresponding actions and relationships) on the other devices in the environment. Users can use learned impact insights to automatically take actions of the devices in the environment.

Analysis Approach. To identify the impact that the most used IoT devices have on the other devices in the environment, we execute the Apriori algorithm on our IoT repository data for each user. We identify the implications of the most used devices on the other devices in the environment as follows:

- We arrange the data in the IoT data repository for each user as shown in Equation 1. In particular, the data matrix rows represent the time-frame and its corresponding device states and columns represent the day of the week.
- We use the data matrix as input to the Apriori algorithm and the algorithm produces the set of implication rules (*i.e.*, patterns). For example, the produced pattern is {Kitchen Light: ON $\implies$ Front Door: Closed}. Right-hand side and Left-hand side of the pattern is known as consequents and antecedents, respectively.
- Based on the most used devices that we identify in RQ1, we filter the produced implication patterns where the antecedents of the expression belong to the most used devices.
- Further, we compute two metrics to measure the strength of the rules, the metrics are Support and Confidence. The Support (S) and Confidence (C) metrics are measured using the Equations 4 and 5, respectively.

$$S\{X1 \implies Y1\} = \frac{\{\#\ of\ times\ a\ pattern\ exists\ in\ the\ given\ dataset\}}{\{Total\ \#\ of\ transactions\ in\ the\ given\ dataset\}} \quad (4)$$

$$C\{X1 \implies Y1\} = \frac{\{\#\ of\ times\ a\ pattern\ exists\ in\ the\ given\ dataset\}}{\{Total\ \#\ of\ transactions\ where\ X1\ exists\ in\ the\ given\ dataset\}} \quad (5)$$

- Finally, we extract the patterns for which the support and confidence values are greater than a half of the transactions using the Confidence Interval threshold (*i.e.*, S{X1 $\implies$ Y1}>=50% and C{X1 $\implies$ Y1}>=50%).

Results. A strong implications patterns are inferred for each user. The identified implication patterns are shown in Table 3. For example, during weekends, User 1 use his/her Hall Light and Home Fan between 8:30 to 11:30am when the Front Door is closed on the majority of the time (*i.e.*, 70%).

We observe that implication patterns with a strong support and confidence metrics can be used in two ways: 1) alert the user about the abnormalities of device behavior in the environment (*e.g.*, intimate the user to shut off the shower when bathroom lights are off to conserve water consumption), or 2) to make smart propagation of certain actions across the devices in the environment (*e.g.*, when the Hall Light and Fan are being used then the system should automatically close the Front door.).

Table 3: The results of the implication patterns produced.

List of devices	Weekdays			Weekends		
	Implication	(S)	(C)	Implication	(S)	(C)
(User 1)						
Hall Light:ON, Home Fan:ON	8:30 to 11:30am			8:30 to 11:30am		
	Outside Light:OFF	58.33%	82.35%	Front Door:Closed	70.00%	87.50%
(User 2)	No strong impliction rules can be infered.					
(User 3)						
Family Room:ON	1 to 1:15am			7:45 to 8am		
	Kitchen: ON	50.00%	83.40%	Foyer: ON	80.00%	80.00%
				1 to 1:15am		
				Kitchen:OFF	80.00%	80.00%
Foyer:ON	7:45pm to 8pm			3 to 3:15am		
	Family Room:OFF	50.00%	85.71%	Bedroom:OFF	60.00%	100.00%
Kitchen: ON				1:15pm to 1:45pm		
				Foyer: ON	60.00%	100.00%
				3:15pm to 4:30pm		
				Foyer: OFF	80.00%	81.00%
(User 4)						
Kitchen:ON, Bedroom L:ON, Bedroom CL: ON	10:45 am to 11:30 am			No strong rules can be inferred.		
	Garage L: ON	73.08%	90.48%			
	Washer: ON	73.08%	90.48%			
	Shower: OFF	73.08%	90.48%			
	Downstairs L: ON	80.77%	100.00%			
	Bathroom: OFF	76.92%	95.24%			
	Bathroom Fan: OFF	76.92%	95.24%			
	Bed LED: OFF	80.77%	100.00%			
	Downstairs L:OFF	80.77%	100.00%			
	Entry Table: OFF	80.77%	100.00%			
	Floor Lamp: OFF	80.77%	100.00%			
	Entryway: OFF	80.77%	100.00%			
	Porch: OFF	80.77%	100.00%			
	Upstairs Light: OFF	80.77%	100.00%			
Track Light Front:ON, Track Light Rear: ON, Kitchen:ON	12pm to 12:15pm			No strong rules can be inferred.		
	Downstairs L: OFF	61.54%	88.89%			
	Bed LED: OFF	69.23%	100.00%			
	Bathroom Fan: OFF	69.23%	100.00%			
	Entry Table: OFF	69.23%	100.00%			
	Floor Lamp: OFF	69.23%	100.00%			
	Entryway: OFF	69.23%	100.00%			
Track Light Front:ON, Kitchen: ON	4pm to 4:15pm			No strong rules can be inferred.		
	Track Light Rear: ON	69.23%	90.00%			
	Nursery Floor: OFF	65.38%	85.00%			
	Garage L: OFF	73.08%	95.00%			
	Shower: OFF	69.23%	90.00%			
	Washer: OFF	73.08%	95.00%			
	Downstairs L: OFF	73.08%	95.00%			
	Bathroom: OFF	76.92%	100.00%			
	Porch: OFF	73.08%	95.00%			
	Bed LED: OFF	76.92%	100.00%			
	Bathroom Fan: OFF	76.92%	100.00%			
	Downstairs Fan: OFF	76.92%	100.00%			
	Entry Table: OFF	76.92%	100.00%			
	Floor Lamp: OFF	76.92%	100.00%			
	Entryway: OFF	76.92%	100.00%			
	Bathroom L: OFF	76.92%	100.00%			
	Upstairs Light: OFF	76.92%	100.00%			

Notes:- (S) and (C) denotes the Support and Confidence of the rule.

*A strong user's behavioural impacts exists among devices which can be used to make smart recommendations (*i.e.,*actions) across the other devices.*

5 THREATS TO VALIDITY

In this section, we discuss the threats to validity of our study through a common guideline [26]:

Internal validity. An internal threat to validity is that we only focus on IoT devices data were posted on the dweet.io publisher-subscription site. The quality of IoT devices data being published might affect the experimental results. To deal with the possible bias, the first author of this work manually verified all the data collected from IoT smart devices to make sure it has the appropriately described data fields.

External validity. An external threat to validity is that we only studied 31 devices from 4 unique users. There are over 2,500 devices available on dweet.io site. Since, our study is focused on discovering knowledge of personalized behaviours in home settings, we analyzed all the available devices and eliminated the devices belonging to industrial settings. Nevertheless, further studies using more IoT devices are welcome.

Construct validity. A construct threat to validity is that the IoT devices data that is used in our study is based on the collection of 30 consecutive days. To remove any possible bias, a longer period of data collection should be performed.

6 RELATED WORK

In this section, we summarize the related work on behavioural knowledge discovery.

Knowledge Discovery. The various techniques from the machine learning, pattern recognition, and artificial intelligence areas can be applied to discover knowledge from machines, sensors, devices, robots, and gadgets data [10]. The research related to the three popular knowledge discovery techniques that are well established in the above disciplines are described below [25]:

(DT1.) **Association Analysis** is a process of uncovering the relationships that exists among the data [23]. The association rules are the models that identify how the data items in a dataset are associated to each other. The association analysis has been used in various research endeavours, such as market analysis (*e.g.*, [5, 14]) and gene classifications (*e.g.*, [3, 7, 19]). Regarding IoT devices data, the association analysis for knowledge discovery is very useful. However, the technique is not been thoroughly explored on IoT devices data for knowledge discovery. In our research, we have used the association rule mining technique to identify the relations among devices in the users' home environment.

(DT2.) **Clustering Analysis** is a process of partitioning a set of analyzed data into subsets [6]. Each subset is represented as a cluster. The data within the same cluster is similar to one another while different among in other clusters. The clustering analysis is a classic knowledge discovery technique. In clustering analysis, various clustering methods are available. In IoT devices data, different clustering methods may yield different varieties of clusters on the same set of analyzed data. For example, clusters, formed on counting the persons inside the garage and the parked vehicles in the garage, are different from clusters formed when trying to count the vehicles parked in the garage by their parked direction. Ortiz *et al.* [16] used the clustering analysis to cluster between the IoT and Social networks to enable the connection of people to the ubiquitous computing devices. Sohn and Lee [21] applied the clustering analysis by ensembling the individual classifiers from two categories of severity, such as property damage and bodily injury in road accidents based on the data collected from devices installed in their city. The clustering analysis is not yet been widely studied among personalized IoT devices. In our research, due to very limited availability of IoT devices for each user the clustering analysis technique was not needed.

(DT3.) **Outlier Analysis** is a process of identifying the data point that is very different from most of the remaining data [1]. The clustering analysis determines the groups of data points that are identical and forms a cluster, whereas outlier analysis identifies the individual data point that are different from the remaining data. Outliers are also commonly referred as abnormalities or anomalies. In IoT devices, the data might not comply with the general actions (*i.e.*,may have abnormalities), such as falsifying the fire alarm at home. The outlier analysis is widely used in research studies, such as the one performed by Elio Masciari which runs an outlier analysis on Radio-Frequency identification (RFID) data streams to identify the tags that are abnormally attached to the objects [15]. Hromic *et al.* [11] used the statistical outlier analysis detection in the real-time sensor data of Internet of Things for events processing using intelligent servers. Kantarci *et al.* [12] used outlier detection analysis for environmental safety by measuring the trustworthiness of data received from cloud-centric IoT. In general, the outliers are detected from the normal data sets so that they can be discarded to keep the study environment pure. The analysis is not yet been vastly studied due to the limited availability of users IoT devices data to public. However, in our research the outlier analysis was not necessary due to the availability of clean IoT devices data.

7 CONCLUSION AND FUTURE WORK

The increasing popularity of the smart devices and its connectivity to the Internet of Things (IoTs) platform created millions of IoT devices. Users use those IoT devices to achieve their personal goals. In this paper, we provide a behavioural extraction engine using Apriori, an association rule mining algorithm, to identify the most used devices by users and to find their relationships with other devices in the environment. In particular, our engine can be used to identify the personalized user behaviour with respect to their devices usage patterns that can be used to alert or make smart recommendations to users in achieving their personal goals.

We conduct a case study from the users' IoT devices usage data collected from dweet.io (*i.e.*, a popular site for publishing and subscribing data of IoT devices) for 4 users. Our results show that, users have on average, 2 IoT devices that they use at specific times and have a relatively small impact across other devices in the environment. Hence, the assimilated users IoT devices behavioural patterns can be used to communicate with users to notify when the abnormality of the device behaviour happens and/or to make smart recommendations; or to propagate actions across devices in the environment automatically, without any user intervention.

In future, we plan to create a larger IoT repository with more types of IoT resources (*e.g.*, industrial IoT devices) and continuous value devices (*e.g.*, thermostat, and heart-rate monitor). Furthermore, we would like to perform our case study on a large scale of user-base and their IoT resources in more environments to achieve more generalized results.

REFERENCES

[1] Charu C Aggarwal. 2015. Outlier analysis. In *Data mining*. Springer, 237–263.

[2] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215. 487–499.

[3] Frank C Arnett, Steven M Edworthy, Daniel A Bloch, Dennis J Mcshane, James F Fries, Norman S Cooper, Louis A Healey, Stephen R Kaplan, Matthew H Liang, Harvinder S Luthra, et al. 1988. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis & Rheumatology* 31, 3 (1988), 315–324.

[4] Peter Pin-Shan Chen. 1976. The entity-relationship modelâĂŤtoward a unified view of data. *ACM Transactions on Database Systems (TODS)* 1, 1 (1976), 9–36.

[5] Yen-Liang Chen, Kwei Tang, Ren-Jie Shen, and Ya-Han Hu. 2005. Market basket analysis in a multiple store environment. *Decision support systems* 40, 2 (2005), 339–354.

[6] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. 1996. *Advances in knowledge discovery and data mining*. Vol. 21. AAAI press Menlo Park.

[7] W James Gauderman. 2003. Candidate gene association analysis for a quantitative trait, using parent-offspring trios. *Genetic epidemiology* 25, 4 (2003), 327–338.

[8] Alexander Gluhak, Srdjan Krco, Michele Nati, Dennis Pfisterer, Nathalie Mitton, and Tahiry Razafindralambo. 2011. A survey on facilities for experimental internet of things research. *IEEE Communications Magazine* 49, 11 (2011).

[9] Lynne Hamill. 2006. Controlling smart devices in the home. *The Information Society* 22, 4 (2006), 241–249.

[10] Jiawei Han, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Elsevier.

[11] H. Hromic, D. Le Phuoc, M. Serrano, A. AntoniÄĞ, I. P. Å¡arko, C. Hayes, and S. Decker. 2015. Real time analysis of sensor data for the Internet of Things by means of clustering and event processing. In *2015 IEEE International Conference on Communications (ICC)*. 685–691. https://doi.org/10.1109/ICC.2015.7248401

[12] B. Kantarci and H. T. Mouftah. 2014. Trustworthy Sensing for Public Safety in Cloud-Centric Internet of Things. *IEEE Internet of Things Journal* 1, 4 (2014), 360–368. https://doi.org/10.1109/JIOT.2014.2337886

[13] Sean Dieter Tebje Kelly, Nagender Kumar Suryadevara, and Subhas Chandra Mukhopadhyay. 2013. Towards the implementation of IoT for environmental condition monitoring in homes. *IEEE Sensors Journal* 13, 10 (2013), 3846–3853.

[14] Bing Liu Wynne Hsu Yiming Ma and Bing Liu. 1998. Integrating classification and association rule mining. In *Proceedings of the 4th.*

[15] Elio Masciari. 2007. A Framework for Outlier Mining in RFID data. In *Database Engineering and Applications Symposium, 2007. IDEAS 2007. 11th International*. IEEE, 263–267.

[16] A. M. Ortiz, D. Hussein, S. Park, S. N. Han, and N. Crespi. 2014. The Cluster Between Internet of Things and Social Networks: Review and Research Challenges. *IEEE Internet of Things Journal* 1, 3 (June 2014), 206–215. https://doi.org/10.1109/JIOT.2014.2318835

[17] Charith Perera, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. 2014. Context aware computing for the internet of things: A survey. *IEEE Communications Surveys & Tutorials* 16 (2014), 414–454.

[18] F John Reh. 2005. Pareto's principle-The 80-20 rule. *BUSINESS CREDIT-NEW YORK THEN COLUMBIA MD-* 107, 7 (2005), 76.

[19] Richa Saxena, Benjamin F Voight, Valeriya Lyssenko, Noël P Burtt, Paul IW de Bakker, Hong Chen, Jeffrey J Roix, Sekar Kathiresan, Joel N Hirschhorn, Mark J Daly, et al. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316, 5829 (2007), 1331–1336.

[20] Anuj Sehgal, Vladislav Perelman, Siarhei Kuryla, and Jurgen Schonwalder. 2012. Management of resource constrained devices in the internet of things. *IEEE Communications Magazine* 50, 12 (2012).

[21] So Young Sohn and Sung Ho Lee. 2003. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. *Safety Science* 41, 1 (2003), 1–14.

[22] Melanie Swan. 2012. Sensor mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensor and Actuator Networks* 1, 3 (2012), 217–253.

[23] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2005. Association analysis: basic concepts and algorithms. *Introduction to data mining* (2005), 327–414.

[24] H Vestberg. 2010. CEO to shareholders: 50 billion connections 2020. (2010).

[25] Qihui Wu, Guoru Ding, Yuhua Xu, Shuo Feng, Zhiyong Du, Jinlong Wang, and Keping Long. 2014. Cognitive internet of things: a new paradigm beyond connection. *IEEE Internet of Things Journal* 1, 2 (2014), 129–143.

[26] Robert K Yin. 2013. *Case study research: Design and methods*. Sage publications.

[27] Michele Zorzi, Alexander Gluhak, Sebastian Lange, and Alessandro Bassi. 2010. From today's intranet of things to a future internet of things: a wireless-and mobility-related view. *IEEE Wireless Communications* 17, 6 (2010).