true. Such hypothetical e2s correlations are not necessarily

true, and have to be verified using event logs (Section 5.3).

Table 1: Part of the adjacency table. A cell marked with

means the corresponding attribute in the column may cor-

5.3

Correlation Mining

relate with the one in the row head. The full table of 73*73

is in our technical report [44]

While there exist many pattern mining methods, few achieve

both good usability and high accuracy in the context of appi-

fied home automation. Supervised mining methods [51,77]

are more accurate but require well annotated datasets or

users' interventions. Unsupervised methods [31, 35, 60, 68]

can be applied to unannotated data, but are less accurate.

Acceleration

CarbonDioxide

Instead of relying on annotated datasets, we propose a

Contact

semantic-based mining method. Semantic information in-

Illuminance

Motion

cludes devices' types and installation locations, which can

Power

be obtained from home automation platforms. Based on this

Presence

information, HAWatcher proposes hypothetical correlations

Humidity

Sound

(in addition to those e2s correlations from smart apps) cor-

Button

Switch

responding to physical channels and user activity channels.

Each hypothetical correlation is then verified independently.

For physical channel correlations, we consider seven phys-

Like other anomaly detection works [35,51,76], we assume

ical properties that are related to many smart home IoT de-

there are no or very few anomalies during the training phase.

vices: illuminance, sound, temperature, humidity, vibration,

5.3.1 Prepossessing Event Logs

power, and air quality. To determine whether two IoT device

attributes may relate via a physical property, we develop an

Prepossessing of event logs is necessary for two reasons: 1)

NLP (Natural Language Processing) based approach. Specifi-

Raw event logs are noisy with repetitive sensor readings. For

cally, for each attribute of an abstract IoT device, we obtain

example, some power meters periodically report similar (but

its description from the SmartThings' developer website [19]

slightly fluctuating) readings. 2) Devices' numeric readings

and parse it into a list of separate words. To objectively eval-

cannot be incorporated into logical calculations. We thus

uate the relatedness between an attribute and a physical

design a preprocessing scheme for redundancy removal and property, we use Google's pre-trained word2vec model [59]

numeric-to-binary conversion.

to calculate the semantic similarity scores between each word

For each device that generates numeric readings, we add up

in the list and the physical property, and use the highest score

its readings from the entire training dataset and calculate its

as the relatedness score between the physical property and

mean u and standard deviation O. Readings that fall outside

the attribute. For each physical property, we select the top

the range [u - are excluded as extreme values

ten attributes with the highest scores, which are considered

(i.e., the three-sigma rule [64]). 3 Then, we apply

the

Jenks

mutually correlated via that physical property.

natural breaks classification algorithm [49]4 to the remaining

readings and classify them as either 'low' or 'high'. Next, for

This way, we are able to find all correlated attribute pairs

and mark them in an adjacency table, part of which is shown

each device's given attribute, we traverse the events and

in Table 1. As SmartThings stipulates 73 attributes [19], the

remove those that do not change the state (e.g., consecutive

table is 73*73. A cell with means that the attributes in its

Fhigh E!!uminance) Now, each two temporally adjacent events about

row head and column head correlate.

the same attribute of a device have opposite values.

While most of the cells are automatically generated, an

5.3.2 Hypothetical Correlation Generation

exception is the switch attribute: as all actuator devices have

the switch attribute, we mark it as correlated with all other

Besides those generated from the smart app channel, hypo-

thetical correlations can be generated from the physical and

attributes. For user activity channel correlations, we use pres-

user activity channels with other semantic information, such

ence and motion as the two special attributes that directly

as device attributes and relations between attributes. We first

reflect users' activities. As a user's activity may affect all

utilize the semantic information to construct a table marking

the attributes, in the adjacency table we mark presence and

motion as correlated with all other attributes.

correlated attribute pairs; then, we fill each pair with de-

vices that have matching attributes to generate hypothetical

For a specific smart home, all attributes of the installed

correlations.

devices are checked against this adjacency table to find pairs

that may correlate. Given a pair of correlated attributes

3 Event exclusion is for training only; the anomaly detection module

a and in the adjacency table, the device A with the at-

does not eliminate events.

4Jenks natural breaks algorithm and K-means algorithm give the same

tribute a, and B with , we generate four hypothetical e2e

results for one-dimension data [38]

correlations (Ed(A),

(Ed(A)