Using Knowledge Graphs for Machine Learning in Smart Home Forecasters

Roderick van der Weerdt[0000−0002−1125−1126]
(Early Stage PhD)

Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
r.p.vander.weerdt@vu.nl

Abstract. Internet of Things (IoT) brings together heterogeneous data from smart devices in smart homes. Smart devices operate within different platforms, but ontologies can be used to create a common middle ground that allows communications between these smart devices outside of those platforms. The data communicated by the smart devices can be used to train the prediction algorithms used in forecasters. This research will first focus on the creation of a mapping to transform IoT data into a knowledge graph than can be used in the common middle ground and investigate the effect of using that IoT knowledge graph data as input for prediction algorithms. Experiments to determine the impact of incorporating other related information in the training of the prediction algorithms will be performed by using external datasources that can be linked to the knowledge graph and by using federated learning over IoT data from other smart homes. Initial results on the transformation mapping of IoT data to an ontology is presented.

Keywords: Internet of Things · Ontology · Data Mapping · Smart Home · Forecasting · Machine Learning · Federated Learning

1 Introduction

Smart devices are on the rise, ranging from dishwashers and televisions to lamps and curtains. We define smart devices as physical objects that are able to communicate, have a unique identifier, with (at least) basic computing power and that may have a sensor [11]. The collection of smart devices and the technologies needed to make them operate is called Internet of Things (IoT). Many different technical platforms are available that allow interactions between devices, but connecting devices functioning on different platforms is often not possible [6], either due to proprietary software, vendor locking, or other implementation choices.

To solve this problem, multiple ontologies [2, 3, 7, 17] have been created to serve as a common middle ground, with the goal of creating interoperability between smart devices. They do this by not simply creating another platform, but by connecting the separate platforms used by all different devices. Devices still operate within their original platform with the common middle ground

allowing communications between the devices by translating the communications to and from their original platforms, whatever they may be, using the central ontology. Our research is done in the context of SAREF (as described in section 2.1) but it is extendable to other ontologies.

Being able to map the smart device data structure into the ontology and then populate it with the smart device data itself creates one knowledge graph that can be used to make a forecaster with combined data from different devices. Using information from related devices has been shown to improve the forecast results of prediction algorithms [16]. Data aggregated from multiple smart home devices will be heterogeneous [13], as it includes different modalities, timescales and data originating from different types of devices (e.g. temperature measurements, camera images or the amount of apples left in the refrigerator). Most machine learning (ML) models are trained on raw data, which is also the case for prediction algorithms used in IoT [9]. Using those models means we would lose the information about the relations between the devices, because we can not use the heterogeneous IoT data as input [19].

The main goal of this research is to investigate the impact of using heterogeneous data from smart devices represented as a knowledge graph to train prediction algorithms used for forecasting in IoT. Before the prediction algorithms can be trained on IoT data, this data first needs to be transformed into the RDF data of a knowledge graph because RDF is capable of handling the heterogeneity of the data.

1.1 Internet of Things

Interoperability through ontologies and learning over knowledge graphs has been done before, the challenges in this research come from applying it to the IoT domain. The two main issues that we address are: distributed knowledge and privacy sensitive data.

distributed knowledge: The goal of the common middle ground is to connect all the smart devices in a smart home, but for a forecaster the data from other smart homes is also useful. Because the data is privacy sensitive it can not be directly shared between different smart homes. Federated learning allows for the sharing of ML model parameters instead of data [10]. These parameters are used to let the other models learn from each other, sharing knowledge about the model and not about the data. Our research will include experiments that use federated learning to increase the learning possibilities of the prediction algorithms.

privacy sensitive data: IoT data from smart devices contains privacy sensitive information about the behaviour of the device's user [6]. For the first experiments we will use open data or refrain from sharing the user data (keep it local in the smart homes environment), but when we use data of multiple users, federated learning will allow us the use of the knowledge in the data without have to share the data itself.

1.2 Interconnect Project

The Interconnect project is a collaboration between 50 partners from 11 European countries with the goal to create "interoperable solutions connecting smart homes, buildings and grids"[1]. One of the pilots part of this project consists of 200 smart homes (apartments with smart devices) that will be built and installed in the Netherlands. Experiments proposed as part of this research will be performed in this setting.

2 State of the Art

The next section will present work related to our research, starting with the ontology that we will use to represent IoT data, the SAREF ontology. The Sections 2.2 through 2.5 present related work on: the transformation process from raw data to knowledge graph data, different ML models that can be used as forecasters, how the knowledge graph can be used directly for the training of ML models and the last subsection describes federated learning and how it relates to IoT and this research.

2.1 SAREF

The Smart Applications REFerence ontology (SAREF) was created for the specific purpose of interoperability between IoT devices from different manufacturers [3]. Figure 1 shows the relations between the main classes used in the ontology. The `saref:Device` class is the central class used in every mapping while the other classes are optional depending on the case under consideration. For example the `saref:Measurement` and `saref:UnitOfMeasure` classes are relevant whenever an observation is mapped, but not when a command to a device is mapped.

SAREF contains some subclasses and instances related to IoT data, such as for example the `saref:Property` subclasses: `saref:Light`, `saref:Motion` and `saref:Temperature`. However more subclasses and instances should be added if needed. To that end SAREF was created to be extendable, allowing for new subclasses or instances to be added or re-used from other ontologies (such as OM-1.8 for units of measurements [14]).

2.2 Creating the Mapping

Creating a mapping to transform the output of the smart home devices into a knowledge graph will allow us to include heterogeneous data and combine data from all smart devices.

Creating this mapping can be done by hand by creating a seperate template for each individual device in RDF (as detailed in Section 6) or it can be done by creating specific mapping rules using a RDF mapping language (RML) [4].

[1] https://interconnectproject.eu/

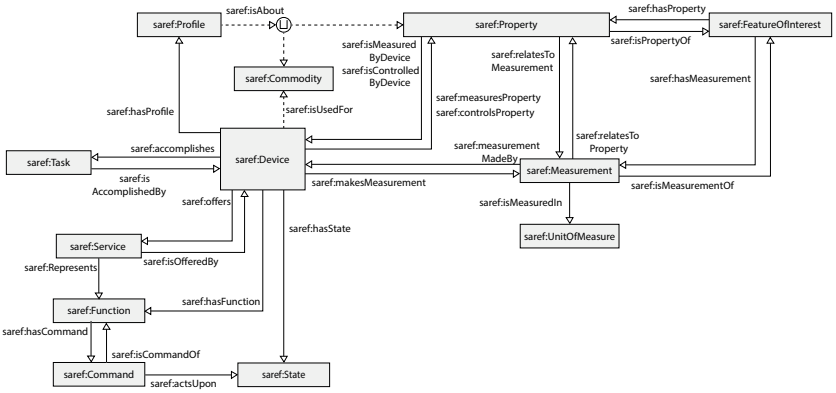

Fig. 1. Overview of the SAREF ontology.[2]

Multiple RMLs have been developed varying in complexity, with the less complex languages being more restricted to the information that is available in the original dataset and the more complex, transformation oriented RDF languages such as D2RML [1], allowing for more elaborate manipulations as, for example, joining data from separate rows.

2.3 Prediction algorithms in IoT

Prediction algorithms are used in smart homes to predict the value of a specific measurement in the future based on its previous measurements and other types of relevant measurements [20]. An example of a prediction algorithm in a smart home is a forecaster that predicts the future temperature of a room based on the current temperature of the room, the temperature of different rooms in the same house and the brightness of the sun on specific walls among other factors [16]. Other possible predictions include category and occurrence time forecasting of daily activity of occupants [21] and energy usage [12].

Multiple forecasters have been created before using machine learning models such as: forward stepwise linear resolution [16], multi-layer perceptron [12], convolutional neural network [21] among other methods.

ML models training data is commonly expected to be raw data (such as the pixels from an image) or a representation of that data (such as vector embeddings representing words). This is also the case with the models presented here, the next section presents methods to transform the knowledge graph without losing its information.

2.4 Learning over Knowledge Graphs

Heterogeneous data is a problem if you want to use it as training data for a ML model [19]. Using a knowledge graph to model that data makes it usable, since

[2] Image taken from: https://saref.etsi.org/core/v3.1.1/#Figure_1

it is not heterogeneous anymore, now it is one knowledge graph that can be used as input. Not many models are able to train directly over knowledge graphs, but there are methods to transform a knowledge graph into data that is accepted by most ML models.

A knowledge graph can be restructured into a table by adding a row for every node-edge-node set in the graph. However, this transformation loses information about the node neighbourhoods around nodes by simplifying to one to one relations.

Graph Neural Networks (GNN) make a knowledge graph processable by creating embedding representations for each node in the graph. Embeddings are vector representations of points in a high dimensional space (vector space). An example of a GGN is Node2vec [5], it trains a model to create embeddings for similar representations, based on neighbouring nodes and the "role" of the node in the graph, placing them closer together in the vector space. The role of a node is based on its position in relation to other nodes (e.g. is it a central node or is it a leaf node). The embeddings can then be used as input for other ML models, because embeddings are strings of numbers, which is a very common input for ML models [9]. Node2vec was used in [8] where a knowledge graph of stock market listed companies was used to create embeddings. The embeddings representing the companies were used to predict the stock price of these companies based on previous results of the company itself and the results of similar companies (because those were the companies closest to it in the vector space).

More elaborate GNNs such as Heterogeneous GNN (HetGNN) [23] consist of multiple models connected end-to-end, allowing for multi-modal data to be embedded by models tailored to the specific kind of data (such as CNNs for images and par2vec for text) before bringing those embeddings together again into one embedding for each node. Zhang et al. demonstrate in their paper how the embeddings can consequently be used with generic ML models for a variety of tasks, such as link prediction, node classification and clustering.

A prediction algorithm for a forecaster could use the embeddings created with a GNN to train a ML model to make predictions based on a set of embeddings. For example, a set of embeddings representing all the measurements made in the previous twelve hours as input and predicted energy consumption in the next twelve hours as output.

2.5 Federated Learning

In federated learning the ML model is trained locally at each data location on the available data, creating a local model. At certain points during the training the parameters of the local models are sent to a central location were they are averaged to create the global model. The parameters from the global are then send back to the local models which continue training with these parameters [10].

The consensus approach does not use a global model, instead the parameters of one smart device hop to the next smart device and the hopped parameters are used to update the local model of the second smart device. Over a number

of communication rounds each device is visited, while the local models are also still trained on the local data [15].

With federated learning we can experiment with larger knowledge bases, using information from other smart homes without having to access the data.

3 Problem Statement and Contributions

The main research question of this work is: *Can we improve the accuracy of prediction algorithms by integrating heterogeneous IoT and Smart energy data and background knowledge?* This research question is broken up into the following subquestions:

RQ1: *Is SAREF an appropriate ontology to model heterogeneous IoT data?* We want to represent as much of the heterogeneous smart device data as possible therefore we want to use a knowledge graph that includes the structure of the data (the ontology) and the data itself (instances in the ontology). We validate SAREF to determine if it is the appropriate ontology to represent all the available information from the smart devices.

RQ2: *Which prediction algorithms are best suited for training on the IoT data knowledge graph?* There are multiple ML models usable as predictions algorithms, but for the forecaster we need to determine which is the best to be used for IoT with a knowledge graph as training data.

RQ3: *Can we improve the accuracy of forecasters by learning over a heterogeneous set of diverse knowledge?* A main advantage of knowledge graphs is that they are linkable with other knowledge graphs, adding more related (for example: outside) data should produce better forecasters.

RQ4: *Can we maintain the accuracy of forecasters with federated learning (over other smart homes)?* In RQ3 we include data from other households in order to increase the data available for training the prediction algorithms, but in practice due to privacy requirements, we can not include the data from other homes directly. Federated learning allows the data from other homes to remain private while making part of their model available to be included in the training of models in other homes. We can formalize this RQ as: given a knowledge graph KG of triples with time bound information, and learning agents that own and see only (overlapping) subsets of this knowledge graph $\mathrm{KG}_i \subseteq \mathrm{KG}$. How can we define a federated learning approach that best predicts target triples (not available in the KG) without agents exchanging information about the triples they see.

3.1 Contributions

The contributions of the research to the state of the art are the following:

- Validation of the SAREF ontology in a realistic and large scale smart home setting.
- A comparison of ML solutions that can train on knowledge graph data.

- A GNN approach that works in an IoT federated learning setting.
- An evaluation of the mapping and forecaster using the ML solutions in a real world case, using the Dutch pilot of the Interconnect project.

4 Research Methodology and Approach

The approach to answering these research questions is detailed in this section for each research question.

RQ1: Before a model can be trained for a forecaster the first step will be the creation of a dataset. To the best of our knowledge there is no IoT data knowledge graph available so it will have to be created by transforming IoT data available in another format to RDF. Non-RDF dataset are available, such as the dataset from [22]. An initial mapping for this transformation has already been performed by hand and is reported on in Section 6. A more generic mapping using RML will be created that is reusable to allow for all the different smart devices data to be transformed. The creation of this mapping allows for the validation of the SAREF ontology, if it able to represent all the data from the smart devices.

RQ2: The second challenge is to create a new forecaster that uses the IoT data knowledge graph as input to train a prediction algorithm. GNN models will be implemented to create predictions based on the data in the IoT data knowledge graph. The resulting forecaster will be tested in a practical setting, as an extended version of the experiment performed in [18], to research the tradeoff between computational resources and accuracy.

This forecasting system will initially be a specific forecaster focusing only on temperature. A second version could be a more general purpose forecaster that is able to collect all the measurements from a smart home and make predictions about what those measurements would look like in the future. This more general purpose forecaster is a more interesting forecaster as it would be able to make the most use of the data collected, using it to make predictions of all the different measurement types.

RQ3: Other datasets with information that is informative for the prediction will also be transformed to RDF to expand the knowledge graph with more information. This can concern datasets about weather forecasts[3], historical weather data[4], information from neighbours or something similar. When the extra data is added to the knowledge graph the prediction algorithm can train a second model with the new knowledge graph to determine the effect the extra data has on the accuracy of the forecaster.

RQ4: We investigate this RQ by using federated learning to collect more information from other smart homes, without directly sharing the data of the smart homes. The Dutch pilot of the Interconnect project with 200 smart homes will be launched in 2022. Federated learning will allow the forecasters in those smart homes to train their ML models on the data from neighbouring homes.

[3] https://data.buienradar.nl/2.0/feed/json
[4] http://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi

5 Evaluation Plan

The evaluations of the results from the methods described in the previous two sections to answer the research questions will be described next.

RQ1: The two implementations described in section 6 and [18] describe how the data from a smart home can be mapped to a knowledge graph, showing that an ontology is capable of mapping relevant data available in a smart home, sufficiently to accomplish a scenario using multiple devices. To provide a more robust answer it will be tested thoroughly in combination with the prediction algorithms used in the other RQs. Competency questions will be defined to determine if the knowledge graph is able to contain all the available information.

RQ2: The IoT data knowledge graph that will be created for RQ1 will be used as input to train state of the art predictors. The prediction task will be to predict the temperature measurement values for the next twelve hours based on the available knowledge of the previous twelve hours. The real values of the temperature measurement values for the next twelve hours will be used as a gold standard for the prediction algorithms to be tested against. Using this gold standard to train multiple prediction algorithm allows us to evaluate the accuracy and efficiency of each model while using the same knowledge graph.

RQ3: Two forecasters will be trained, one on the knowledge graph as used in RQ2 and one on the new extended knowledge graph. The same gold standard created for RQ2 will be used to calculate the accuracy of the new model. A significance test will then be used to show if the accuracy of model trained on the extended knowledge graph significantly differs from the accuracy of the model trained only on the "original" knowledge graph.

RQ4: Two models will be used for the evaluation of RQ4. One model trained using a prediction algorithm that uses federated learning to include data from neighboring houses. And a second model that is again trained with the "original" knowledge graph from RQ2. Applying a significance test on the resulting accuracy of both models will provide a clear answer to the research question.

6 Intermediate Results

To create an initial dataset of RDF data, a mapping for the output of nine different smart devices was created by hand [18]. These mappings were created in collaboration with domain experts knowledgeable about the devices and experts knowledgeable about the SAREF ontology.

To answer RQ1 an experiment with SAREF in a practical setting was created that used the Knowledge Engine [18]. The Knowledge Engine is a custom built interoperability framework, built by a partner from the Interconnect project, that allows multiple smart devices to communicate using RDF data. This was achieved by adding to every smart device a smart connector, a single board computer with a script capable of communicating directly with "its" smart device in whatever format it was programmed. It sends the data in RDF to the other smart connectors, which in turn interprets it and send it to their smart device in the formats that it accepts.

We set up an experiment that recreated a scenario from a smart home setting.One smart connector was connected to a thermometer, one smart connector was connected to a smart thermostat and one smart connector was connected to a smart heater. Through this communication the smart thermometer was able to send its current temperature measurement, which was received by the smart thermostat that compared it to an internal setting (controlled with buttons) and based on this either send a `saref:OnState` or `saref:OffState` to the smart heater. All of these messages were possible to express using SAREF.

7 Conclusions

This research plan shows how we intend to answer the research question: *Can we improve the accuracy of prediction algorithms by integrating heterogeneous IoT and Smart energy data and background knowledge?*

After this first year of research the first research question has been partially answered by the two implementations of SAREF described in [18], showing how an ontology can be used to create a knowledge graph of IoT data. But more research will have to determine whether a knowledge graph can hold all the information required for the prediction algorithms that will be used for the other research questions.

When this research is completed it will introduce a new way of combining the information from connected IoT devices, resulting in more accurate forecasters and leading to more efficient smart homes.

Acknowledgements. This work is part of the Interconnect project (interconnectproject.eu/) which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 857237.

I would like to thank Victor de Boer, Laura Daniele, Frank van Harmelen, Ronald Siebes and Steffen Staab for their guidance and feedback.

References

1. Chortaras, A., Stamou, G.: Mapping diverse data to rdf in practice. In: International Semantic Web Conference. pp. 441–457. Springer (2018)
2. Compton, M., Barnaghi, P., Bermudez, L., GarcíA-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., et al.: The SSN Ontology of the W3C Semantic Sensor Network Incubator Group. Journal of Web Semantics **17**, 25–32 (2012)
3. Daniele, L., den Hartog, F., Roes, J.: Created in Close Interaction with the Industry: the Smart Appliances REFerence (SAREF) Ontology. In: International Workshop Formal Ontologies Meet Industries. pp. 100–112. Springer (2015)
4. Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: Rml: a generic language for integrated rdf mappings of heterogeneous data. In: Ldow (2014)
5. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 855–864 (2016)

6. Hsu, C.L., Lin, J.C.C.: An Empirical Examination of Consumer Adoption of Internet of Things Services: Network Externalities and Concern for Information Privacy Perspectives. Computers in Human Behavior **62**, 516–527 (2016)
7. Janowicz, K., Haller, A., Cox, S.J., Le Phuoc, D., Lefrançois, M.: SOSA: a Lightweight Ontology for Sensors, Observations, Samples, and Actuators. Journal of Web Semantics **56**, 1–10 (2019)
8. Long, J., Chen, Z., He, W., Wu, T., Ren, J.: An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in chinese stock exchange market. Applied Soft Computing **91**, 106205 (2020)
9. Mahdavinejad, M.S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., Sheth, A.P.: Machine learning for internet of things data analysis: A survey. Digital Communications and Networks **4**(3), 161–175 (2018)
10. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. pp. 1273–1282. PMLR (2017)
11. Miorandi, D., Sicari, S., De Pellegrini, F., Chlamtac, I.: Internet of things: Vision, applications and research challenges. Ad hoc networks **10**(7), 1497–1516 (2012)
12. Nawaz, A., Hafeez, G., Khan, I., Jan, K.U., Li, H., Khan, S.A., Wadud, Z.: An Intelligent Integrated Approach for Efficient Demand Side Management With Forecaster and Advanced Metering Infrastructure Frameworks in Smart Grid. IEEE Access **8**, 132551–132581 (2020)
13. Qin, Y., Sheng, Q.Z., Falkner, N.J., Dustdar, S., Wang, H., Vasilakos, A.V.: When things matter: A survey on data-centric internet of things. Journal of Network and Computer Applications **64**, 137–153 (2016)
14. Rijgersberg, H., van Assem, M., Top, J.: Ontology of Units of Measure and Related Concepts. Semantic Web **4**(1), 3–13 (2013)
15. Savazzi, S., Nicoli, M., Rampa, V.: Federated learning with cooperating devices: A consensus approach for massive iot networks. IEEE Internet of Things Journal **7**(5), 4641–4654 (2020)
16. Spencer, B., Al-Obeidat, F.: Temperature Forecasts with Stable Accuracy in a Smart Home. Procedia Computer Science **83**, 726–733 (2016)
17. W3C: Web of Things (WoT) Thing Description (2020), https://www.w3.org/TR/2020/REC-wot-thing-description-20200409/
18. van der Weerdt, R., de Boer, V., Daniele, L., Nouwt, B.: Validating saref in a smart home environment. Metadata and Semantic Research **1355**, 35–46 (2021). https://doi.org/10.1007/978-3-030-71903-6_4
19. Wilcke, X., Bloem, P., De Boer, V.: The knowledge graph as the default data model for learning on heterogeneous knowledge. Data Science **1**(1-2), 39–57 (2017)
20. Wu, S., Rendall, J.B., Smith, M.J., Zhu, S., Xu, J., Wang, H., Yang, Q., Qin, P.: Survey on Prediction Algorithms in Smart Homes. IEEE Internet of Things Journal **4**(3), 636–644 (2017)
21. Yang, H., Gong, S., Liu, Y., Lin, Z., Qu, Y.: A multi-task learning model for daily activity forecast in smart home. Sensors **20**(7), 1933 (2020)
22. Zamora-Martínez, F., Romeu, P., Botella-Rocamora, P., Pardo, J.: On-line learning of indoor temperature forecasting models towards energy efficiency. Energy and Buildings **83**, 162–172 (2014)
23. Zhang, C., Song, D., Huang, C., Swami, A., Chawla, N.V.: Heterogeneous graph neural network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 793–803 (2019)