Sampling IoT Knowledge Graph to Enhance Computational Efficiency

BSc Thesis *(Afstudeerscriptie)*

written by

Raihan Karim Ishmam
(born Septemer 12th, 2003 in Bangladesh)

under the supervision of **Dr. Victor de Boer** and **Mr. Roderick van der Weerdt**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

Bachelor in Artificial Intelligence
Intelligent Systems Track

at the *Vrije Universiteit Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
July 21st, 2023	Dr. Ronald Siebes (second reader)

Abstract

In recent years, IoT knowledge graphs have emerged as a focal point in the scientific community, playing a pivotal role in addressing the challenges of data integration. Not only has it attracted widespread attention for its convenient semantic interoperability, but has also unlocked immense potential for data analytics to serve in the IoT domain. One issue is that, because of the heterogeneous and longitudinal format of data collected by devices, IOT KGs may be very large. This can lead to challenges in loading and querying the knowledge graphs, resulting in inefficiency in the process. Sampling techniques offer a practical way to tackle the size issue of IoT knowledge graphs, thereby enabling efficient analysis. This study investigates the sampling of IoT KGs, reducing them to a manageable size while maintaining good computational properties to answer relevant user queries. We investigate this in the form of a case study, where we work with data from multiple heterogeneous devices in a Dutch office building, collected in the context of the Interconnect project. Our algorithm uses semantic filters based on relational properties to sample desired parts of the KG. The relational properties are hosted by the SAREF ontology, making the algorithm extendable to other SAREF-based KGs. We evaluated the method by querying the sampled knowledge graph for the end-user-provided usage scenarios. The results show that the sampled knowledge graph retains completeness with regard to our user-end objectives, despite being reduced to a smaller size.

Keywords— IoT, Knowledge Graphs, Interconnect Project, SAREF

Contents

Chapter 1

Introduction

Linked data have become an integral part of the data science field, enabling the integration and interlinking of diverse data sources to a semantic web of shared data [9]. As part of this paradigm, knowledge graphs have emerged as a powerful representation and organization mechanism for connected data [15]. Knowledge graphs capture the relationships and semantic connections between entities, forming a network of interlinked entities. The capabilities of knowledge graphs extend to facilitate advanced data exploration and analysis, as they can link data from multiple domains. In the domain of the Internet of Things (IoT), knowledge graphs play a crucial role in harnessing the potential of interconnected devices and sensors through the data they collect [5].

The field of Internet of Things (IoT) has witnessed a surge in interest and research focus in recent years, with IoT knowledge graphs emerging as a crucial tool within the scientific community [21]. IoT knowledge graphs enable a holistic and interconnected view of IoT data, fostering enhanced understanding and decision-making. Knowledge graphs play a pivotal role in addressing the complex challenges associated with data integration in the IoT domain. By facilitating semantic interoperability and unlocking the vast potential of data analytics, IoT knowledge graphs have garnered significant attention.

1.1 Problem Statement

However, a notable challenge in the realm of IoT knowledge graphs is the large size of these graphs, which arises from the heterogeneous and longitudinal nature of the collected data. As IoT devices generate diverse data formats over time owing to their wide range of devices, the resulting knowledge graphs can become remarkably large [15]. Furthermore, the granularity combined with the longitudinal data creates a size concern as the devices track the data history via time points [21], adding to the graph size.

Chapter 4

Evaluation

As expected, the sampling algorithm successfully sampled 365 files. Despite variations among the sizes of each file, the knowledge graph was reduced to 25% of the size of the source knowledge graph. Therefore, we observed a close relationship between the devices and the size proportions between the source and the sampled knowledge graph. In this section, we evaluate the utility of the resulting knowledge graph. To assess whether the knowledge graph maintained the computational properties required to answer the user questions, we ran SPARQL queries to determine whether the knowledge graph had the extracted entities *(see Section 3.2)* required to answer the questions. The resulting knowledge graph was loaded into GraphDB, where the SPARQL queries were run on it.

4.1 Temperature agaisnt Time

```
PREFIX ic: <https://interconnectproject.eu/example/>
PREFIX saref: <https://saref.etsi.org/core/>

SELECT DISTINCT (?value AS ?Temparature)
                (?dateTime AS ?Time)
WHERE {

    ic:device_urn:Device:SmartThings:9a6268d1-e0cd-484d
    -95cd-42737fa4bce0
                saref:makesMeasurement ?measmeasurement.
    ?measurement saref:relatesToProperty ?property.
    ?property a saref:Temperature.
    ?measurement saref:hasvalue ?value;
                saref:hasTimestamp ?dateTime.
```

Listing 4.1: SPARQL query for temperature and time

A thermostat device was randomly chosen to obtain an insight into how the temperature recorded by the thermostat varied over time. *Listing 4.1* shows the query used to extract the temperature and time data of the device. *Figure 4.1* depicts a visualization of the temperature against time from the query results plotted using *Matplotlib*[1] [22]. By observing the patterns and fluctuations depicted in the visualization, we can gain insights into the impact of certain times or activities during those times on the temperature. *For a monthly overview of the temperature data, refer to the appendix.*

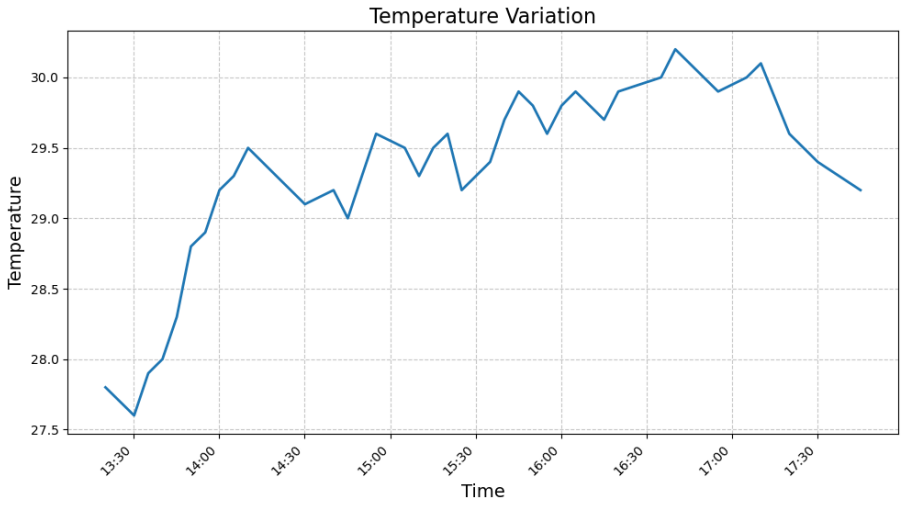

Figure 4.1: Temperature Variation Hourly

4.2 Windows Status against Time

```
PREFIX ic: <https://interconnectproject.eu/example/>
PREFIX saref: <https://saref.etsi.org/core/>

SELECT DISTINCT (IF(?value = 1, 'Open', 'Close') AS ?Status)
                (SUBSTR(STR(?dateTime), 1, 10) AS ?Date)
                (SUBSTR(STR(?dateTime), 12, 8) AS ?Time)

WHERE {
```

[1] https://matplotlib.org/

5.1 Future Work

Our study has laid the baseline for future research to attempt to answer the user questions from VideoLab, using our sampled knowledge graph. The prospect could to extended to address more scenarios involving properties not touched upon in our paper. In future research, it would be worthwhile to explore ways to extend the applicability of the algorithm to other standard ontologies and to adapt it to handle diverse data representations. Furthermore, efforts to enhance the flexibility of the algorithm and overcome limitations in differentiating devices or data sources could lead to more comprehensive sampling techniques. By addressing these aspects, the potential of the algorithm can be broadened to encompass a wider array of IoT domains and scenarios, thereby contributing to the advancement of efficient IoT knowledge graph analysis and utilization.

This study makes no definitive claim regarding the full usability of consumer-grade devices for sampled Knowledge Graphs. Nevertheless, by undertaking this research on a device of this grade, we have taken a small yet significant step towards advancing the field of knowledge graph sampling and enabling research to be conducted effectively on consumer-grade platforms. This contribution brings us closer to harnessing the potential of everyday devices and opens new possibilities for expanding the utilization of Knowledge Graphs in practical applications.

Acknowledgement

I wish to extend my sincere appreciation to my supervisor, Victor de Boer, for his constant guidance and support throughout this thesis project. Victor's invaluable insights, continuous guidance, and constructive feedback greatly shaped the structure and direction of this paper. I am also deeply grateful to Roderick van der Weerdt for his pivotal role in this project. Roderick's commitment to facilitating access to essential resources substantially contributed to the realization of this research.