Evaluating the Effect of Semantic Enrichment on Entity Embeddings of IoT Knowledge Graphs

Roderick van der Weerdt, Victor de Boer, Laura Daniele, Ronald Siebes and Frank van Harmelen
May 28th, 2023
Introduction

Setting the scene

• Ongoing effort to collect IoT data
  ‣ For example in IoT Knowledge Graphs
• There are benefits to learning over KGs
• Can we learn directly over the IoT KG?
  ‣ Or do IoT KGs require changes?

Evaluating the Effect of Semantic Enrichment on Entity Embeddings of IoT Knowledge Graphs
Introduction

Setting the scene

• Ongoing effort to collect IoT data
  ‣ For example in IoT Knowledge Graphs
• There are benefits to learning over KGs
• Can we learn directly over the IoT KG?
  ‣ Or do IoT KGs require changes?

What is the effect of semantically enriching an IoT KG, based on the quality of entity embeddings learned from it?

Evaluating the Effect of Semantic Enrichment on Entity Embeddings of IoT Knowledge Graphs
Introduction

What is an IoT KG?

• Represent IoT measurement data

• "Wide" Graph

• Many measurements
Introduction

What is Semantic Enrichment?

- Making implicit information explicitly available
Introduction

What are Entity Embeddings?
Introduction
Why do we need Entity Embeddings?

Evaluating the Effect of Semantic Enrichment on Entity Embeddings of IoT Knowledge Graphs
Introduction

Research question

What is the effect of semantically enriching an IoT KG, based on the quality of entity embeddings learned from it?
What is the effect of semantically enriching an IoT KG, based on the quality of entity embeddings learned from it?
Experimental Pipeline
Overview

Step I
- Basic Graph
- Semantic Enrichment
- Enriched Graph

Step II
- Embedding Model
- Entity Embeddings

Step III
- Classification Task
- Accuracy

Accuracy
compare

Entity Embeddings
Entity Embeddings
Experimental Pipeline Overview

- Step 1: Semantic Enrichment
- Step 2: Embedding
- Step 3: Evaluation
Experimental Pipeline
Semantic Enrichment step

- Rounded value
- Sequence links
- Timestamp
Experimental Pipeline

Embedding step

- RDF2vec
- Walk length of 2
- 25 walks per entity
Experimental Pipeline
Embedding step

- RDF2vec
- Walk length of 2
- 25 walks per entity
Experimental Pipeline

Evaluation step - Classification Task

• Separated all timepoints based on outside temperature

• Labeled warmest half “warm” & coldest half “cold”

• Trained a MLP to classify timepoints based on label
**Experimental Pipeline**

**Evaluation step - Accuracy comparison**

- Experiment is performed with both the original graph and enriched graph

- Accuracies are compared
Experimental Pipeline

Dataset

- OPSD Household dataset
- Hourly measurements
- 8133 timepoints (±11 months)
- Made into a IoT KG

<table>
<thead>
<tr>
<th></th>
<th>device types</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>grid import</td>
</tr>
<tr>
<td>res1</td>
<td>1</td>
</tr>
<tr>
<td>res2</td>
<td>1</td>
</tr>
<tr>
<td>res3</td>
<td>1</td>
</tr>
<tr>
<td>res4</td>
<td>1</td>
</tr>
<tr>
<td>res5</td>
<td>1</td>
</tr>
<tr>
<td>res6</td>
<td>1</td>
</tr>
<tr>
<td>total</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 2 shows the distributions of the different device types over each separate residence.

The dataset contains measurements taken over a five years duration, but not every device recorded measurements for the entire period. In order to have a complete dataset we chose to extract a subset of ten months where all devices had recorded measurements, by removing only two devices (the freezer from residence 2, and the grid export from residence 6), if we would have included these the measurements would only be available for two months.

The final manipulation of the data was transforming the energy consumption measurement from its original value of accumulated consumption from the startpoint, to accumulated consumption over the last hour. This manipulation was performed to ensure that the measurement values in the graph would be recurring, which would not be the case for accumulated measurement values because those would only increase.

The final graph represents 8133 timestamp entities linking to measurements from 37 devices from ten device types, spread out over six residences. Three different versions of this graph were created in order to be able to distinguish between the effects of adding more devices from within the same home, and adding devices from other homes. The following shorthand is used to refer to different compositions of the graph:

- res1dev1: this graph uses only measurements of one device. This is the heat pump from residence 4. This graph contains 89477 triples.
- res1devA: this graph uses all measurements from all devices in one home, in this case, all devices from residence 4. This graph contains 715,765 triples.
- resAdevA: this graph uses all measurements from all devices in all available homes that are available. It contains 3,220,912 triples.

The IoT KGs can be found at: https://github.com/RoderickvanderWeerdt/SAREFized-OPSD-household-graph

- OPSD Household dataset
- Hourly measurements
- 8133 timepoints (±11 months)
- Made into a IoT KG

https://data.open-power-system-data.org/household_data/2020-04-15
### Experimental Pipeline

**Dataset**

- **res1dev1:**
  1 device from 1 residence

- **res1devA**
  All devices from 1 residence

- **resAdevA**
  All devices from All residences

---

Table 2

<table>
<thead>
<tr>
<th>device types</th>
<th>residence id</th>
<th>grid import</th>
<th>grid export</th>
<th>pv</th>
<th>dishwasher</th>
<th>ev</th>
<th>refrigerator</th>
<th>freezer</th>
<th>heat pump</th>
<th>washing machine</th>
<th>circulation pump</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>res1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>res2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>res3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>res4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>res5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>res6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td></td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>39</td>
</tr>
</tbody>
</table>

The IoT KGs can be found at: [https://github.com/RoderickvanderWeerdt/SAREFized-OPSD-household-graph](https://github.com/RoderickvanderWeerdt/SAREFized-OPSD-household-graph)
Results

Average accuracies of the classifiers

• Enriched outperforms basic every time

• More “useless” devices have a negative impact

• More “useful” devices have a positive impact
Discussion

Ongoing Research

• Using a different evaluation task
  ‣ Value prediction

• Using a different embedding method
  ‣ GCN

• Using different datasets
  ‣ Pecan street (American consumption data)
Ongoing Research

Average accuracies of the classifiers - RDF2vec & GCN

GCN

RDF2vec
Ongoing Research
Average accuracies of the VALUE PREDICTOR - RDF2vec & GCN

**GCN**

- res1dev1: 6.0
- res1devA: 6.5
- resAdevA: 7.0

**RDF2vec**

- res1dev1: 5.5
- res1devA: 6.0
- resAdevA: 6.5
Conclusion

What is the effect of semantically enriching an IoT KG, based on the quality of entity embeddings learned from it?

- In this setting, enriching an IoT KG has a **positive** effect on the quality of entity embeddings learned from it.
What is the effect of semantically enriching an IoT KG, based on the quality of entity embeddings learned from it?

- In this setting, enriching an IoT KG has a **positive** effect on the quality of entity embeddings learned from it.

Thank you for listening