Augmenting Weather Sensor data with Remote Knowledge Graphs

Jiantao Wu\textsuperscript{1,2}  Fabrizio Orlandi\textsuperscript{1,3}  Declan O’Sullivan\textsuperscript{1,3}  Soumyabrata Dev\textsuperscript{1,2}

July 19, IGARSS 2022

\textsuperscript{1}The ADAPT SFI Research Centre, Ireland

\textsuperscript{2}School of Computer Science, University College Dublin

\textsuperscript{3}School of Computer Science and Statistics, Trinity College Dublin, Ireland
1. Motivation & Objective

2. Methodology

3. Experiments

4. Future Work
Motivation

In the data-driven meteorology area, the **variety** and the **amount** of data can often significantly affect the performance of analytical models\(^1\), *E.g.* machine learning, deep learning...

- **Traditional data acquisition**
  - Data bulks, fragmentation
  - Schema mismatch
  - Varied formats, *E.g.* CSV, JSON, HTML...

- **Knowledge graph (KG)**
  - Consistent format throughout Web, *I.e.* RDF
  - Linked Data techniques - interlinked KGs over HTTP
  - Semantics - data can be manipulated closer to how humans think!

---

\(^1\) Tarek AlSkaif et al. “A systematic analysis of meteorological variables for PV output power estimation”. In: *Renewable Energy* (2020)
To explore the remote KGs in its strength of acquiring data in an interoperable manner and how it can efficiently augment the sensor data for improved analytical models.

We have

- applied SPARQL language to gain a powerful access to multi-KG for meteorological data, and
- explored the effectiveness of multi-KG in rainfall detection improvement.

In the spirit of reproducible research, all the source code is available at https://github.com/futaoo/multiKG.
1. Motivation & Objective

2. Methodology

3. Experiments

4. Future Work
Knowledge graphs preparation

Figure: NOAA climate KG (Left) and PurpleAir air quality KG (Right)

Note:

- The schema has not to be fixed in a KG
- Schemas, E.g. *sosa*, can be shared among different KGs
- the query protocol (i.e. **SPARQL**) is identical for Two KGs
Federating SPARQL queries

BASE <http://jresearch.ucd.ie/climate-kg/>
PREFIX ca_prop: <http://jresearch.ucd.ie/climate-kg/ca/property/>

SELECT ?prcp ?humidity ?date
WHERE { #KG1
  ?obs ca_prop:sourceStation <resource/station/GHCND:EI000003969>;
    sosa:hasSimpleResult ?prcp;
    sosa:hasResult/ca_prop:withDataType <resource/datatype/PRCP>;
    sosa:resultTime ?date .
}
SELECT ?humidity ?date
WHERE { #KG2
  SERVICE <http://jresearch.ucd.ie/kg/air-pollutants/sparql> {
    ?atm_obs sosa:madeBySensor <purpleair/sensor?id=26695>;
    sosa:hasSimpleResult ?humidity;
    sosa:observedProperty <http://jresearch.ucd.ie/climate-kg/purpleair/Temperature>;
    sosa:resultTime ?date.
  }\}} FILTER (YEAR(?date)=2019)}

- Query multiple KGs at once
  - with shared semantics
  - “acquire (data) as you go”

- Time alignment for multiple time series
  - to prepare multivariate time series
  - to increase the predicting accuracy!
1. Motivation & Objective

2. Methodology

3. Experiments

4. Future Work
Enhanced rainfall detection with Multi-KG

Task Definition:

The probability threshold is set to be 0.5 above which we claim the rainfall will happen on a future day. Let $\mathbf{x}^t$, $\mathbf{x} \in \mathbb{R}^N$ be a $n$-dimensional vector at time step $t$. $\mathbf{x}^t = (x_1^t, x_2^t, ..., x_N^t)$ where $x_n^t$ denotes a KG1’s meteorological variable or KG2’s atmospheric variable at time step $t$. $y \in \{0, 1\}$ denotes the binary label and $y = 1$ (PRCP > 0) means that rainfall occurs. Hence a formal definition of rainfall detection on time step $t$ based on the observations of previous $k$ time steps can be formulated as Equation 1:

$$P^t(y = 1|(\mathbf{x}^{t-1}, \mathbf{x}^{t-2}, ..., \mathbf{x}^{t-k}))$$  (1)
Motivation & Objective

Methodology

Experiments

Future Work

Enhanced rainfall detection with Multi-KG

Results:

Table: Training data constructed for 2-step multivariate time series classification

<table>
<thead>
<tr>
<th></th>
<th>Humidity (t-2)</th>
<th>PRCP (t-2)</th>
<th>TAVG (t-2)</th>
<th>Humidity (t-1)</th>
<th>PRCP (t-1)</th>
<th>TAVG (t-1)</th>
<th>Rain (t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>57.95</td>
<td>1.2</td>
<td>13.3</td>
<td>51.37</td>
<td>0.0</td>
<td>14.9</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>51.37</td>
<td>0.0</td>
<td>14.9</td>
<td>51.49</td>
<td>9.3</td>
<td>11.5</td>
<td>yes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>366</td>
<td>37.19</td>
<td>0.0</td>
<td>14.9</td>
<td>39.92</td>
<td>0.0</td>
<td>14.6</td>
<td>no</td>
</tr>
</tbody>
</table>

Table: Performance evaluation of rainfall detection on two datasets: (a) KG1’s meteorological data, and (b) dataset (a) enhanced with KG2’s atmospheric variables.

<table>
<thead>
<tr>
<th>Models</th>
<th>Datasets</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>KG1</td>
<td>0.66</td>
<td>0.63</td>
<td>0.65</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>KG1 + KG2</td>
<td>0.71</td>
<td>0.77</td>
<td>0.74</td>
<td>0.68</td>
</tr>
<tr>
<td>SVC</td>
<td>KG1</td>
<td>0.65</td>
<td>0.73</td>
<td>0.69</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>KG1 + KG2</td>
<td>0.70</td>
<td>0.85</td>
<td>0.76</td>
<td>0.69</td>
</tr>
<tr>
<td>KNN</td>
<td>KG1</td>
<td>0.69</td>
<td>0.51</td>
<td>0.59</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>KG1 + KG2</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
<td>0.64</td>
</tr>
</tbody>
</table>
1. Motivation & Objective

2. Methodology

3. Experiments

4. Future Work
Future Work

We will work on:

- Developing an advanced pipeline with a graphical user interface to assist users in augmenting KG data for machine learning applications, and
- Incorporating more data domains for climate research, such as remote sensing data, tourism data, and so on.