# Towards a Semantic Outlier Detection Framework in Wireless Sensor Networks

Extended Abstract

Iker Esnaola-Gonzalez
IK4-TEKNIKER
Iñaki Goenaga 5
Eibar, Spain 20600
iker.esnaola@tekniker.es

Jesús Bermúdez
University of the Basque Country (UPV/EHU)
Paseo Manuel Lardizabal 1
Donostia-San Sebastian, Spain 20018
jesus.bermudez@ehu.eus

Izaskun Fernández
IK4-TEKNIKER
Iñaki Goenaga 5
Eibar, Spain 20600
izaskun.fernandez@tekniker.es

Santiago Fernández
IK4-TEKNIKER
Iñaki Goenaga 5
Eibar, Spain 20600
santiago.fernandez@tekniker.es

Aitor Arnaiz
IK4-TEKNIKER
Iñaki Goenaga 5
Eibar, Spain 20600
aitor.arnaiz@tekniker.es

## ABSTRACT

Outlier detection in the preprocessing phase of Knowledge Discovery in Databases (KDD) processes has been a widely researched topic for many years. However, identifying the potential outlier cause still remains an unsolved challenge even though it could be very helpful for determining what actions to take after detecting it. Furthermore, conventional outlier detection methods might still overlook outliers in certain complex contexts. In this article, Semantic Technologies are used to contribute overcoming these problems by proposing the SemOD (Semantic Outlier Detection) Framework. This framework guides the data-scientist towards the detection of certain types of outliers in WSNs (Wireless Sensor Network). Feasibility of the approach has been tested in outdoor temperature sensors and results show that the proposed approach is generic enough to apply it to different sensors, even improving the accuracy, specificity and sensitivity of outlier detection as well as spotting their potential cause.

## KEYWORDS

Outlier Detection, Semantic Technologies, Wireless Sensor Network, Knowledge Discovery in Databases

**ACM Reference format:**
Iker Esnaola-Gonzalez, Jesús Bermúdez, Izaskun Fernández, Santiago Fernández, and Aitor Arnaiz. 2017. Towards a Semantic Outlier Detection Framework in Wireless Sensor Networks. In *Proceedings of SEMANTiCS 2017, Amsterdam, The Netherlands, September 11-14, 2017,* 8 pages.
https://doi.org/10.475/123_4


*SEMANTiCS 2017, September 11-14, 2017, Amsterdam, The Netherlands*

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

## 1 INTRODUCTION

Today's real-world datasets are highly susceptible to noisy, missing, and inconsistent data due to their typically large size and their likely origin from multiple, heterogeneous sources [7]. Low-quality data can complicate the knowledge extraction process, leading to low-quality mining results or even inaccurate conclusions. That is why it is important to ensure data quality in KDD processes. There are several data preprocessing techniques to increase data quality, which are not mutually exclusive and may be applied together (e.g. filtering and missing data treatments).

Outliers are data objects that stand out amongst other data objects and do not conform to the expected behaviour in a dataset [10]. The process of finding these data objects is known as Outlier Detection and it is an essential task for a wide range of domains including intrusion detection for cyber-security, fault detection in safety critical systems, fraud detection for credit cards and data monitoring in WSNs (Wireless Sensor Networks).

Outliers can occur for various reasons and understanding their provenance helps to determine what actions to take after detecting them. In some cases the aim might be to isolate the outlier and act on it (e.g. fraud detection) while in others, outliers are filtered out to avoid inaccurate results (e.g. data analytics). However, identifying the potential cause of outliers still remains an unsolved challenge in most cases: it is not always straightforward and it may become an arduous task. There are also challenging scenarios where a data object may be considered an outlier in one context (e.g. 40°C measurement is an outlier for a winter day in the north of Spain), but not an outlier in a different context (e.g. 40°C measurement is not an outlier for a summer day in the south of Spain).

In this article, the SemOD (Semantic Outlier Detection) Framework is presented. This framework is based on Semantic Technologies to guide the data-scientist through the Outlier Detection process in WSNs. An ontology-based reasoning process infers the circumstances that can affect the sensor's measurements, and the data-scientist is provided with methods comprising purporsely defined steps and a set of resources to detect outliers generated in such contexts. That is, each method detects the potential provenance of

outliers, which provides valuable information for the decision making. Results show that the proposed approach is generic enough to apply it to different sensors and it might detect outliers that classic outlier detection methods might overlook when context is insufficient [11] or rather complicated to be modelled without Semantic Technologies.

Throughout the paper, Section 2 reviews related work in outlier detection methods, WSNs and Semantic Technologies. SemOD Framework is presented in Section 3 as well as its main modules. Section 4 shows experiments performed to test the approach and Section 5 evaluates and discusses obtained results. Finally in Section 6 conclusions and future work are presented.

## 2 RELATED WORK

Outlier detection has been a widely researched topic for many years and there has been an abundance of work from statistics, geometry, machine learning, database, and data mining communities.

Model-based approaches create statistical distribution models assuming that data objects not following the model are outliers. Distance-based outlier detection approaches study the distance of a data object to its nearest neighbours and use it to measure its outlierness. A typical distance-based method is the k-NN, which uses the distance to the nearest k neighbours to determine whether a data object is an outlier or not. Density-based methods compare the density of a data object in a neighbourhood and the ones with lower density are considered outliers. Outlier detection methods based on clustering assume that normal data objects belong to big dense clusters, whereas the ones belonging to small or sparse clusters are outliers. Further information regarding conventional outlier detection methods can be found in [1] and [8].

One of the fields in which Outlier Detection is relevant is WSNs, where several factors make them prone to outliers due to their particular requirements, dynamic nature and resource limitation. Apart from these factors, WSNs are also context dependent, so conventional outlier detection techniques may not be directly applicable [16] or can have their results skewed. [2] makes use of correlations between sensors to detect and classify outliers in WSNs. Outliers are classified into erroneous data or caused by an actual event occurred in the physical world[1].

In the last years, advantages of Semantic Technologies for data understanding as well as for the data mining process itself have been highlighted in [9] and [13]. Furthermore, many approaches have proposed the use of semantically annotated data to enhance different KDD phases. A detailed and extended survey on Semantic Technologies within the KDD process can be found in [14], where it is stated that while many impressive results can be achieved already today, the full potential of Semantic Technologies for KDD is still to be unlocked.

Regarding the data quality field, it has been tackled specially in [3], [4] and [5], where data quality problems in Semantic Web data (e.g. missing and illegal values or functional dependency violations) are identified by means of data validation and SPIN (SPARQL Inferencing Notation) rules. A review of the existing data quality works based on ontologies for the health domain is shown in [12].

[1]This classification is not comparable to the one proposed by the SemOD Framework (see Section 3), where outliers are classified according to their potential cause.

Although a well researched topic, outlier detection has not received sufficient attention from the Semantic Web Community. In [15] a domain ontology has been used to support the outlier detection based on a statistical method. In [6] segment outliers and unusual events are detected in WSNs combining statistical analysis and domain expert knowledge captured via ontology and semantic inference rules. This approach determines whether the sensor collects suspicious data or not by calculating its similarity with neighbours. To the extent of our knowledge, this proposal is one of the few works where Semantic Technologies have a direct role in outlier detection methods. However, it may not be applicable to isolated nodes where there are no nearby sensors to compare its similarity. Furthermore, the identification of the potential cause of outliers is not tackled in this approach.

We believe that the role of Semantic Technologies in Outlier Detection could be more important and could have a prominent impact not only improving the outlier detection, but most importantly in the assistance of data-scientists during this process and spotting the potential cause of outliers. The proposed SemOD Framework focuses on contributing in these issues.

## 3 SEMOD FRAMEWORK

The SemOD Framework guides the data-scientist through the Outlier Detection process in WSNs. This framework is based on an ontology (EEPSA Ontology) containing domain and expert-knowledge to identify circumstances that make these sensors susceptible to errors. Each of these circumstances has been assigned a method (SemOD Method) in which constraints that describe outliers are generated. These constraints are generated in a (semi)automatic way following purposely defined steps and using a set of resources. These have been designed by experts in a way that no previous knowledge regarding the domain or Semantic Technologies are required to take advantage of them. Data-scientist is then assisted to make use of these constraints to generate a SPARQL query (SemOD Query) which retrieves sensor measurements that are presumably outliers because of being measured under a certain circumstance.

The framework is intended for novice data-scientist as well as data-scientists non-experts in the domain, since outliers are detected in a (semi)automatic way and with no previous knowledge required. Additionally, it can also be a valuable tool for expert data-scientists, who many times may overlook potential causes of outliers when trying to detect them in WSNs.

The SemOD Framework is composed of three main modules: the EEPSA Ontology, the SemOD Method and the SemOD Query.

### 3.1 EEPSA Ontology

The EEPSA (Energy Efficiency Prediction Semantic Assistant) Ontology[2] aims to capture all the necessary knowledge related to sensing and actuating devices, as well as their corresponding observations and actuations. Furthermore, it contains expert knowledge regarding the energy efficiency, buildings and WSN domains. In the SemOD Framework, the EEPSA Ontology's use has a twofold purpose. On the one hand, it is used for semantic annotation. On the other, it contains essential information for the generation of constraints and queries in the upcoming modules of the framework.

[2]https://w3id.org/eepsa

First of all, information about sensors, measurements, and the context surrounding them have to be semantically annotated with terms contained in the EEPSA ontology. Sensors are represented with class *sosa:Sensor* which is reused from the SSN Ontology[3]. Furthermore, Measurements4EEPSA module[4] is imported to represent spatially located things (e.g. *m3-lite:HumiditySensor*), orientations (e.g. *m3-lite:hasDirection* property and individuals *m4eepsa:northEastOrientation*) and units of measurements (e.g. *m4eepsa:WaterConductivity*). Measurements made by sensors are represented with class *sosa:Observation* and properties measured by sensors are represented with the property *sosa:observedProperty* and suclasses of *ssn:Property*. A class *eepsa:Outlier* is defined as a subclass of *sosa:Observation*, which will be populated (as well as it subclasses like *eepsa:OutlierCausedByRain*) with outliers detected by SemOD Queries.

Once the required information is semantically annotated and due to the OWL axioms within the EEPSA Ontology, a reasoner can infer circumstances that make sensors susceptible to suffer from outliers. That is, when sensors are under those circumstances, their measured values are likely to be outliers. As previously stated, several circumstances make WSN and sensors prone to errors. For example, an indoor temperature sensor located in a poorly isolated external wall can be conditioned by external meteorological conditions such as wind speed or solar radiation. When exposed to rain, a wet outdoor sensor will not measure the same humidity as a dry sensor due to the evaporation of water from its surface. A sensor placed next to a light bulb might have its illuminance measurements affected when the bulb is on. A sensor might not make accurate measurements if power supply levels are not enough. These are just some of the circumstances affecting sensors that are gathered in the EEPSA ontology.

So, this first module of the SemOD Framework allows the data-scientist to identify circumstances that make sensors susceptible to outliers.

## 3.2 SemOD Method: Outdoor Temperature Sensor Affected by Solar Radiation

In order to detect outliers caused by each of those circumstances, the corresponding SemOD Method must be applied. SemOD methods provide data-scientists with purposely defined steps and a set of resources towards the (semi)automatic generation of a SemOD Query to detect outliers caused by a certain circumstance. These resources and steps have been designed by experts in such a way that no previous knowledge regarding the domain or Semantic Technologies are required to take advantage of them.

The SemOD Method that detects outliers measured by outdoor temperature sensors has been developed as a starting point of the research. This SemOD Method focuses on detecting outliers caused when sensors receive direct solar radiation. When this happens, sensors tend to get hot and measure much higher temperatures than real ones. It is composed of three steps that are shown in Figure 1.

[3]http://www.w3.org/ns/ssn/
[4]https://w3id.org/measurements4eepsa

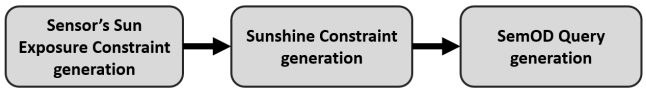


**Figure 1: SemOD Method steps for detecting outliers caused by solar radiation in an outdoor temperature sensor.**

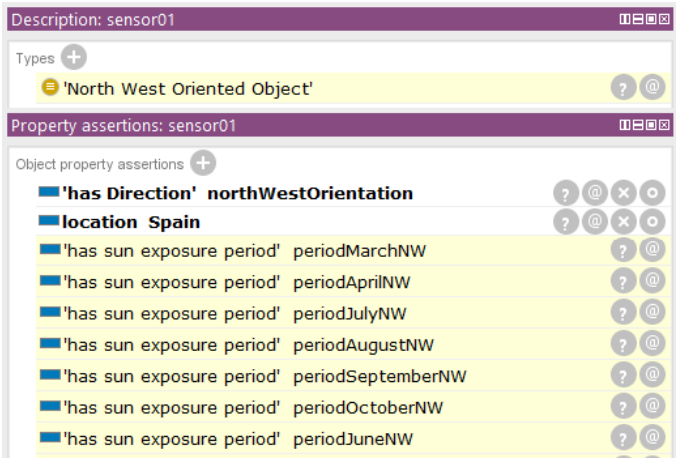


**Figure 2: Excerpt of individual *sensor01*.**

*3.2.1 1st step: Sensor's Sun Exposure Constraint generation.* For a measurement to be affected by solar radiation, this SemOD Method specifies that two conditions must happen at the same time. On the one hand, sensor in charge of measuring has to be placed in a position where, during measurement time receives direct solar radiation. On the other, there must be no obstacles such as clouds on the sun beam lights' way to the sensor.

According to experts, time spans when an object might be exposed to direct solar radiation depends mainly on its location, orientation and the time of the year. When the orientation and location of an object are semantically annotated, thanks to EEPSA ontology's OWL axioms, a reasoner can infer the periods in which the object might be exposed to the sun. Each of these periods is described by means of *eepsa:startingTime*, *eepsa:endingTime* and *eepsa:hasMonth* properties. For example, a sensor *sensor01* located in Spain and oriented towards the Northwest is inferred to be an individual of class *eepsa:NorthWestOrientedObject*. As a part of its definition, any individual belonging to this class will have a *eepsa:hasSunExposurePeriod* property with values such as *periodFebruaryNW* as it can be seen in Figure 2. Therefore, retrieving attribute values of *eepsa:periodFebruaryNW*, it can be concluded that on February, *sensor01* (see Figure 3) is exposed to sun between 18:00 and 19:00.

Keeping sun exposure times generic for any location is not a feasible task since these vary depending on the location. Those values expressed here are acceptable for locations in Spain. However, even in Spain there might be places where these values might not be completely accurate. In case periods' starting and ending times need to be refined, data-scientist is always free to do so in the

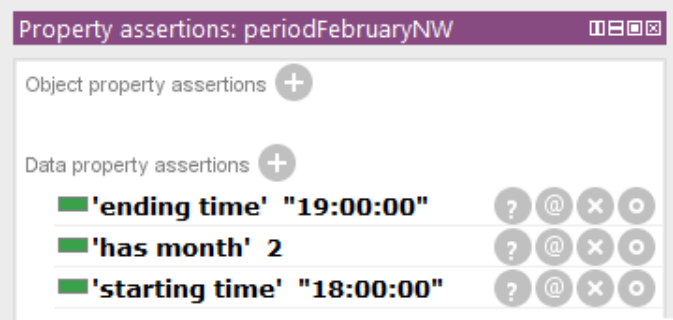


**Figure 3: Excerpt of individual *periodFebruaryNW*.**

sensor instance during the next stage of the SemOD Framework (SemOD Query Execution).

The SemOD Method offers a constraint pattern describing an object's sun exposure times as presented in Listing 1. This constraint pattern is composed of a month (integer) and two times in *hh:mm:ss* format, so that it retrieves objects that happen during the month and between the two times. In order to instantiate this pattern, wild cards *MONTH_VALUE*, *STARTING_TIME* and *ENDING_TIME* need to be replaced with values contained in the ontology assertions.

```
(month(?date) = MONTH_VALUE &&
?time >= STARTING_TIME && ?time <= ENDING_TIME)
```

**Listing 1: Constraint pattern describing an object's sun exposure times.**

In order to retrieve sun exposure times of an object, the SemOD Method proposes the SPARQL query pattern shown in Listing 2. Wild card OBJECT needs to be replaced with the corresponding object's URI.

```
SELECT *
WHERE
{       OBJECT eepsa:hasSunExposurePeriod ?period.
        ?period eepsa:startingTime ?startingTime;
                eepsa:endingTime ?endingTime;
                eepsa:hasMonth ?monthValue. }
```

**Listing 2: SPARQL query pattern retrieving an object's sun exposure times.**

Values retrieved by the SPARQL query are used to instantiate the constraint pattern in Listing 1. This has to be instantiated as many times as exposure periods the object has. Each instantiation of the constraint pattern has to be linked with the next one using the OR (||) operator. For example, a fragment of the instantiation produced by *sensor01* would look like Listing 3.

```
(month(?date) = 2 &&  ?time >= xsd:time("18:00:00")
 && ?time <= xsd:time("19:00:00") ) ||
(month(?date) = 3 &&  ?time >= xsd:time("17:00:00")
 && ?time <= xsd:time("20:00:00") ) || ...
```

**Listing 3: Constraint pattern describing *sensor01*'s sun exposure times.**

These constraints only concern sensor's sun exposure times. As previously stated, it is also necessary that during these periods solar radiation hits the sensor.

*3.2.2 2nd step: Sunshine Constraint generation.* In order to resolve whether the sensor receives direct solar radiation or not, the SemOD Method requires information coming from one of these two properties: illuminance or solar irradiance. That means that it is possible to determine whether there is sunny weather or not using measurements of any of these two properties. Experts have established that if threshold values of 70W/m2 for solar irradiance and 15,000lx for illuminance are exceeded it means that it is sunshine.

SemOD Method defines two sources of information to retrieve values for these properties. Firstly, the sensor that is measuring outdoor temperature and secondly, Open Data. Sensor information is considered to be more adequate to create the constraints. So, when information coming from these two sources is available, it is preferable to use the one coming from the sensor itself. Most of times, chances are that Open Data will provide information for a nearby location, but not for the exact sensor location, which can skew results.

The query pattern shown in Listing 4 can be used to determine whether the sensor observes additional properties such as solar irradiance or illuminance. Wild card SENSOR has to be replaced with the sensor URI being analysed (in this case, *sensor01*'s URI). Likewise, this query could be used to retrieve all sensors measuring Solar Irradiance or Illuminance, if wild card SENSOR is replaced with a query variable such as ?sensor and ASK is replaced with SELECT *.

```
ASK
WHERE {
 { SENSOR ssn:hasSubSystem ?sensor
   ?sensor sosa:observedProperty
                eepsa:SolarIrradiance }
UNION
 { SENSOR ssn:hasSubSystem ?sensor
   ?sensor sosa:observedProperty
                m3-lite:Illuminance   }   }
```

**Listing 4: SPARQL query to determine whether a sensor measures Solar Irradiance or Illuminance properties.**

If a sensor measures any of these properties, it can be derived whether sun hits the sensor or not, using the threshold values that experts have previously established. This information is used to complete previous constraints. For example, *sensor01* measures temperature and solar irradiance, so that the constraints created in Listing 3 could be completed adding new information regarding sunshine, resulting in Listing 5.

```
(month(?date) = 2 &&  ?time >= xsd:time("18:00:00")
 && ?time <= xsd:time("19:00:00")
 && xsd:integer(?solarIrradianceVal)
         > xsd:integer(70) )
|| ...
```

**Listing 5: SPARQL query to determine whether a sensor measures Solar Irradiance or Illuminance properties.**

In case a sensor neither has illuminance nor solar radiation measuring capabilities, SemOD Method recommends to retrieve this information from Open Data. In our use cases, Euskalmet[5] (Basque Meteorology agency) weather stations have been chosen as source of information. These stations' information is available in different formats in Open Data Euskadi[6] (Basque Open Data portal). Some of them were semantically annotated as part of a previous research and are publicly available via SPARQL endpoint[7]. After executing the GeoSPARQL query[8] shown in Listing 6 in the mentioned endpoint, weather stations that measure solar irradiance or illuminance are retrieved. By default weather stations are ordered by distance, so LAT and LON wild cards need to be replaced with the latitude and longitude of the sensor itself. In case this information is unknown, replacing these wild cards with coordinates of the building or city in which sensor is deployed would also work.

```
SELECT DISTINCT ?stationID ?stationName
 (bif:st_distance(
 (bif:st_point(xsd:float(?lat), xsd:float(?lon))),
 (bif:st_point(xsd:float(LAT), xsd:float(LON)))))
  AS ?distanceToBuilding
FROM <http://tekniker.es/euskalmetStations>
WHERE {
  ?weatherStation rdf:type eepsa:WeatherStation;
        foaf:name ?stationName;
        geo:latitude ?lat;
        geo:longitude ?lon;
        dc:Identifier ?stationID;
        s4bldg:contains ?sensor.
  ?sensor ssn:hasSubSystem ?sensorComponent.
  ?sensorComponent sosa:observes ?property.
FILTER (
  ?property = eepsa:SolarIrradiance
  || ?property = m3-lite:Illuminance
)}
ORDER BY ?distanceToBuilding
```

**Listing 6: GeoSPARQL query retrieving Weather Stations measuring Solar Irradiance or Illuminance.**

Once the most adequate weather station is selected, its hourly measurements can be retrieved from the Open Data Euskadi portal as well. Since weather stations data may have heterogeneous structures depending on the agency they are controlled by, it is very difficult to propose a generic process applicable to all of them. For the Euskalmet weather stations' measurements, we have developed a process that extracts data, annotates it semantically and stores it in a RDF Store, from which it can later be extracted using SPARQL queries.

*3.2.3 3rd step: SemOD Query generation.* The resulting constraints from the previous phase (Listing 5) have to replace the wild card PREVIOUSLY_GENERATED_CONSTRAINTS in the FILTER clause of predefined SemOD Query pattern shown in Listing 7. These constraints also need to be casted into their corresponding data types. Furthermore, the graph where the query is going to be performed needs to be specified in the FROM clause, replacing the RDF_GRAPH wild card. Wild cards PROPERTY and UNIT_OF_MEASUREMENT need also to be specified with the property and unit URI used to derive sun information (i.e. solar irradiance or illuminance).

```
CONSTRUCT {?obs1 rdf:type
   eepsa:OutlierCausedBySolarRadiation}
FROM <RDF_GRAPH>
WHERE {
 ?sensor1 sosa:observedProperty
        m3-lite:Temperature.
 ?sensor2 sosa:observedProperty PROPERTY;
   eepsa:hasUnitOfMeasure UNIT_OF_MEASUREMENT.
 ?obs1 sosa:isObservedBy ?sensor1;
   eepsa:obsTime ?time;
   eepsa:obsDate ?date;
 ?obs2 sosa:isObservedBy ?sensor2;
   eepsa:obsTime ?time;
   eepsa:obsDate ?date;
   sosa: hasSimpleResult ?illuminanceVal.
FILTER(
 PREVIOUSLY_GENERATED_CONSTRAINTS  )  }
```

**Listing 7: SemOD Query pattern for detecting outliers caused by solar radiation.**

Finally, the SemOD Query is generated and ready to be executed. Listing 8 shows a snippet of a SemOD Query that has been generated for *sensor01* case[9]. As previously stated, proposed SemOD Query is intended to be generic enough and adequate for every location in Spain. However, we are aware that values used in constraints might need to be fine-tuned in some cases. Data-analyst is free to do so in this step of the SemOD Method.

```
CONSTRUCT {?obs1 rdf:type
   eepsa:OutlierCausedBySolarRadiation}
FROM <myGRAPH>
WHERE {
 ?sensor1 sosa:observedProperty
        m3-lite:Temperature.
 ?sensor2 sosa:observedProperty
    m3-lite:Illuminance;
   eepsa:hasUnitOfMeasure m3-lite:Lux.
 ?obs1 sosa:isObservedBy ?sensor1;
   eepsa:obsTime ?time;
   eepsa:obsDate ?date;
 ?obs2 sosa:isObservedBy ?sensor2;
   eepsa:obsTime ?time;
   eepsa:obsDate ?date;
   sosa:hasSimpleResult ?illuminanceVal.
FILTER(
  (month(?date) = 2 && ?time > xsd:time("18:00:00")
  && ?time < xsd:time("19:00:00")
  && xsd:integer(?illuminanceVal) )
         > xsd:integer(15000) )
  || ... )  }
```

**Listing 8: SemOD Query generated to detect temperature outliers caused by solar radiation.**

[5]http://www.euskalmet.euskadi.eus
[6]http://opendata.euskadi.eus/
[7]http://193.144.237.227:8890/sparql
[8]PREFIX clauses are not included to save space. The complete query is available at http://193.144.237.227:8890/DAV/home/dba/NearbyEuskalmetStationsSPARQL.txt
[9]The complete SemOD Query is available at http://193.144.237.227:8890/DAV/home/dba/SemODQueryExample.txt

When executed in the next module, this query will retrieve temperature measurements likely to be outliers because of the sensor being hit by solar radiation.

### 3.3 SemOD Query execution

Generated SemOD Query has to be executed over the sensor measurements' dataset to detect measurements suspected to be outliers because of receiving direct solar radiation. This measurements will also be classified as individuals of class *eepsa:OutlierCausedBySolar Radiation*. Depending on the needs of the use case, it is up to the data-scientist what to do with these detected outliers (e.g. remove them from the dataset, analyse them, etc.).

This query is generated in a (semi)automatic manner and with no previous knowledge required for the data-scientist. Furthermore, not only does detect outliers, but also classifies them according to their potential cause.

The SemOD Framework is an ongoing work expected to be completed in following stages of the research. A knowledge elicitation process of circumstances that make sensors susceptible to outliers have already been done and this knowledge is soon to be captured in the EEPSA Ontology. Each of these circumstances will have a corresponding SemOD Method, containing a set of steps and resources designed by experts so that data-scientists can use them to detect outliers caused by those circumstances.

## 4 EXPERIMENTS

Feasibility of the proposed SemOD Framework was tested in different sensors located in the IK4-TEKNIKER building (from now on referred as Tekniker) which is located in Eibar (Basque Country, Spain). These were developed as a part of the European FP-7 Tibucon project[10] and they observe temperature, humidity and illuminance properties with a periodicity of 5 minutes. All measurements are stored in a PostgreSQL database.

In order to determine if an outlier detected by an outlier detection technique is an actual outlier or not, a reference dataset has been used to make comparisons with. Namely, this dataset is composed of temperature measurements made by a Euskalmet weather station located about 6km away from Tekniker and with a similar environment. This station is equipped with a Rotronic sensor to measure temperature and all hourly measurements are available in the Basque Open Data portal. On average Tekniker sensor measurements have a deviation of 2.3°C compared with the stations measurements. Keeping this in mind, a temperature difference of 5°C was set as a threshold to determine whether a temperature measured in Tekniker is an actual outlier or not. That means that a temperature measured in Tekniker differing in more than 5°C compared with the reference one, will be considered as an actual outlier.

Table 1 summarizes the features of the three sensors in which experiments were performed. The measurements column determines the number of measurements available after sampling sensor data with a hourly frequency, the time interval defines the time spans in which sensors made the measurements, and the actual outliers column determines the number of measurements with more than 5°C of difference comparing with the reference dataset.

[10]http://www.tibucon.eu

Before using the SemOD Framework, a statistical outlier detection technique was applied to obtain some baseline results. SemOD enabled query's detected outliers were later evaluated by comparing with baseline results. After testing different algorithms offered by Rapidminer, best results were obtained with the Detect Outlier (Densities) operator.

First of all, in order to apply the SemOD Framework Tekniker building, sensors themselves and measurements were semantically annotated with the EEPSA ontology. For measurements, an ETL (Extract Transform Load) process was defined, which extracted data from the original database, annotated them semantically according to the EEPSA ontology and stored them in a Virtuoso Server 07.20.3217 version, running on an Ubuntu 14.04 server[11]. With all required data semantically annotated, a HermiT 1.3.8.413 reasoner infers circumstances that make sensors susceptible to outliers. Amongst those circumstances, focus was placed on the outliers caused by sensors receiving direct solar radiation (*SolarRadiationCircumstance*). SemOD Method proposed in section 3.2 was applied to generate a SemOD Query to detect outliers potentially caused by this circumstance.

Apart from temperature, Tibucon sensors also measure illuminance, so information about the sun can be derived from sensor itself. However, the SemOD Method's variant proposing the use of Open Data was also tested in the same three sensors. For this case, solar radiation measurements coming from a Euskalmet weather station[12] were semantically annotated based on the EEPSA Ontology using the JENA framework. Obtained results were not as accurate as expected, since the adequate weather station could not be found for our cases.

## 5 EVALUATION AND RESULTS DISCUSSION

In this section, results obtained from experiments described in the previous section are discussed and evaluated. Table 2 summarizes these results.

SemOD Framework enabled outlier detection techniques (referred to as SemOD in this section) slightly improve accuracies compared with the baseline outlier detection techniques, except for the case of the T7 sensor. However, the remarkable outcome is that potential provenance of outliers detected by SemOD is known, while other classical outlier detection techniques give no meaningful insight to outlier causes. As explained later on the case of sensor T23, this provenance provides valuable information for the decision-making process.

The dataset that comprises T17 sensor's measurements is part of a predictive problem project. Therefore, analysis of the applied outlier detection techniques has focused on specific needs of this problem: having a high specificity (detection of actual outliers, a.k.a. recall) while not neglecting the sensitivity (normal data not being mistakenly classified as outliers). Both SemOD and baseline techniques have high specificity, being SemOD the one with the highest (99.7% against baseline's 99.6%). As for sensitivity, a big leap can be observed from 26% of outliers detected by the baseline,

[11]This RDF Store is private because it is considered to contain sensitive data. A sample of raw sensor data is available at
http://193.144.237.227:8890/DAV/home/dba/DataSample.csv

[12]This weather stations is not the same stations that was used as a reference to determine the detected actual outliers.

**Table 1: Tibucon sensors features and measurements summary.**

| Sensor ID | Orientation | Measurements | Actual Outliers | Time intervals |
|---|---|---|---|---|
| T17 | Northwest | 4,074 | 768 | 02/2016 - 07/2016 |
| T7 | Southwest | 5,226 | 547 | 02/2016 - 11/2016 |
| T23 | Northwest | 1,540 | 73 | 02/2017 - 05/2017 |

**Table 2: Summary of obtained results after applying baseline and SemOD outlier detection techniques.**

| Sensor ID | Applied Technique | Accuracy | Specificity | Sensitivity |
|---|---|---|---|---|
| T17 | Baseline | 85.7% | 99.6% | 26% |
| T17 | SemOD | 86.8% | 99.7% | 31% |
| T7 | Baseline | 91.7% | 99.4% | 35.7% |
| T7 | SemOD | 89.6% | 99.9% | 15.7% |
| T23 | Baseline | 94.3% | 98.7% | 17.8% |
| T23 | SemOD | 95.1% | 100% | 11% |

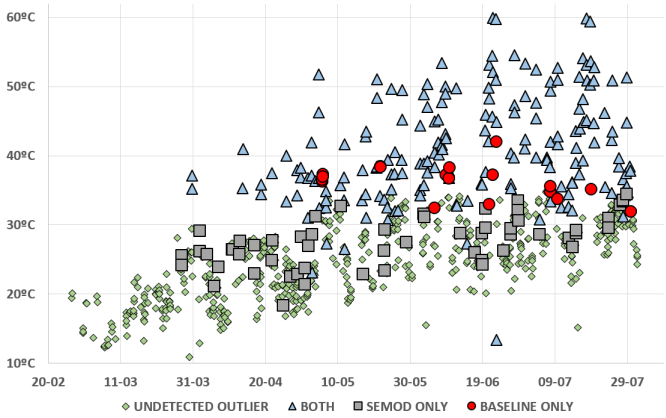


**Figure 4: Actual outliers detected by different techniques**

against the 31% detected by SemOD. Figure 4 plots actual outliers partitioned in four kinds: the ones detected by the SemOD, the ones detected by the baseline technique, the ones detected by both of them, and the ones undetected by any of them.

Out of the existing 768 actual outliers in the dataset, both SemOD and baseline techniques detect the same actual outliers in more than 75% of cases. It is remarkable that baseline technique only detects outliers with high temperatures, while SemOD also detects outliers with lower temperatures. For example, SemOD detects an 18°C outlier measured the 24th of April at 18:00 - a temperature value that may not seem exceptional and could be considered as an ordinary temperature. However, this temperature was measured while sensor T17 was being hit by the sun, leading to the measurement of a much higher temperature in contrast with the 11.7°C measurement made by the reference weather station. The SemOD detected this outlier in a straightforward way supported by the semantic annotation of the context, which is an essential part of the SemOD Framework.

Looking at the chart, it is also noticeable that most actual outliers are overlooked by both outlier detection techniques applied. It has to be clear that the applied SemOD Method only focused on the detection of outliers caused by sensor receiving solar radiation. Therefore, it can be concluded that many undetected outliers may be caused by other circumstances.

Analysing the results of T7 sensor, SemOD obtains a higher specificity than the baseline technique (99.9% against 99.4%). Nevertheless, the number of detected actual outliers dropped from baseline's 35.7% to SemOD's 15.7%. Results have to be interpreted by taking into consideration that SemOD only looks for outliers caused by direct solar radiation, whereas baseline looks for all of them. Consequently, it is reasonable that the baseline technique detects more actual outliers in general.

Looking at Table 1 obtained results for T7 might be somewhat unexpected. The sensor is oriented in a direction more exposed to sun than others (in our location, southwest oriented objects have longer exposure times than northwest-oriented objects) and it contains more measurements. In spite of these factors, less actual outliers occur comparing with sensor T17. What happens in this case is that T7 sensor is better shielded from the sun thanks to Tekniker's design (the Southwest-oriented side of the building is protected from the sun most of the year by a window overhang). Since T7 is actually less exposed to sun, less actual outliers are caused by solar radiation and that is why SemOD technique detects a low percentage (15.7%) of actual outliers. The rest of them presumably happen by other causes.

It is very difficult to compare performance of different outlier detection techniques in T23 sensor's dataset because it is very unbalanced (less than 5% of measurements are actual outliers). SemOD improves accuracy when comparing with the baseline results. However, the most noteworthy thing is that sensor T23 was placed after applying SemOD Framework on T17 sensor and discovering that many outliers were caused by direct solar radiation. T23 and T17 are placed few meters away from each other but T23 is strategically placed so that it is sheltered from sunshine the majority of the day. A direct comparison of both sensors' datasets cannot be made because measurements for the same periods are not available. But T23 suffered from 3 times less outliers than T17 did during

the same period of time the previous year. This supports our claim that spotting the potential provenance of outliers can aid in the decision-making.

## 6 CONCLUSIONS

In this article, the SemOD Framework is presented. This framework is based on Semantic Technologies to guide the data-scientist through the Outlier Detection process in WSNs. An ontology-based reasoning process infers the circumstances that can affect sensor's measurements, and data-scientist is provided with methods to detect outliers generated in such contexts. Due to having context and data themselves annotated, it has been proved that outliers caused in certain contexts are detected in a straightforward way (e.g. in T17 sensor), without the need of creating additional contextual attributes. Furthermore, each method detects the potential provenance of outliers, which provides valuable information for the decision making such as determining how to prevent or act on them. An illustrative example is sensor T23, which was strategically placed after detecting that a relevant circumstance behind T17 sensor's outliers was the direct solar radiation on the sensor.

The proposed SemOD Framework's genericity has been proved by applying it in different sensors. Moreover, this framework provides data-scientists with purposely defined steps and with a set of resources towards the (semi)automatic generation of a SemOD Query to detect outliers. These resources and steps have been designed by experts in such a way that no previous knowledge regarding the domain or Semantic Technologies are required to take advantage of them.

### 6.1 Future Work

SemOD Framework is an ongoing work expected to be completed in future stages of the research. Direct solar radiation is a circumstance making temperature sensors susceptible to outliers but not the only one. EEPSA Ontology is being extended to cover circumstances that make sensors susceptible to outliers. For each of these circumstances, a SemOD Method comprising a set of steps and resources need to be designed so that data-scientists may use them in a (semi)automatized way and with no previous knowledge required.

Notwithstanding the promising results obtained, SemOD Framework was only tested in a few sensors in the same location. It is expected to apply this approach to sensors located in a different building in a different location Bilbao (Basque Country, Spain).

After analysing obtained results, it seems like the SemOD Query detecting outliers caused by solar radiation might oversimplify context at times, so that not all outliers generated by this circumstance were detected. For example, when a sensor gets heated for being exposed to solar radiation for a long time but in the moment of the measurement sun gets covered by a cloud, the generated query might not detect that measurement as outlier. In order to avoid this situations, further semantic annotation is planned to be employed, such as the amount of time of a sensor being exposed to solar radiation.

Finally, a set of interfaces are planned to be developed in order to facilitate the interaction with the SemOD Framework.

## ACKNOWLEDGMENTS

This work is partly supported by the project BID3A3 (Big Data para RIS3 2017) from the Basque Government (ELKARTEK2016) under grant agreed reference KK-2017/00056. This work is also supported by FEDER/TIN2016-78011-C4-2-R and FEDER/TIN2013-46238-C4-1-R. We thank Euskalmet (Basque Meteorology Agency) and Zuzenean (Basque Open Data Portal) for their assistance.

## REFERENCES

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 15.
[2] Asmaa Fawzy, Hoda MO Mokhtar, and Osman Hegazy. 2013. Outliers detection and classification in wireless sensor networks. *Egyptian Informatics Journal* 14, 2 (2013), 157–164.
[3] Christian Fürber. 2015. *Data quality management with semantic technologies*. Springer.
[4] Christian Fürber and Martin Hepp. 2010. Using semantic web resources for data quality management. In *International Conference on Knowledge Engineering and Knowledge Management*. Springer, 211–225.
[5] Christian Fürber and Martin Hepp. 2010. Using SPARQL and SPIN for data quality management on the semantic web. In *International Conference on Business Information Systems*. Springer, 35–46.
[6] Lianli Gao, Michael Bruenig, and Jane Hunter. 2014. Semantic-based detection of segment outliers and unusual events for wireless sensor networks. *arXiv preprint arXiv:1411.2188* (2014).
[7] Jiawei Han, Micheline Kamber, and Jian Pei. 2011. *Data mining: concepts and techniques*. Elsevier.
[8] Victoria J. Hodge and Jim Austin. 2004. A survey of outlier detection methodologies. *Artificial Intelligence Review* 22, 2 (2004), 85–126.
[9] Krzysztof Janowicz, Frank Van Harmelen, James A. Hendler, and Pascal Hitzler. 2014. Why the data train needs semantic rails. *AI Magazine* (2014).
[10] Vijay Kotu and Bala Deshpande. 2014. *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Morgan Kaufmann.
[11] Jiongqian Liang and Srinivasan Parthasarathy. 2016. Robust Contextual Outlier Detection: Where Context Meets Sparsity. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2167–2172.
[12] Siaw-Teng Liaw, Alireza Rahimi, Pradeep Ray, Jane Taggart, S. Dennis, Simon de Lusignan, B. Jalaludin, AET Yeo, and Amir Talaei-Khoei. 2013. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. *International journal of medical informatics* 82, 1 (2013), 10–24.
[13] Qudamah K. Quboa and Mohamad Saraee. 2013. A state-of-the-art survey on semantic web mining. *Intelligent Information Management* 5 (2013), 10–17.
[14] Petar Ristoski and Heiko Paulheim. 2016. Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web* (2016).
[15] Yongheng Wang and Shenghong Yang. 2010. Outlier detection from massive short documents using domain ontology. In *Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on*, Vol. 3. IEEE, 558–562.
[16] Yang Zhang, Nirvana Meratnia, and Paul Havinga. 2010. Outlier detection techniques for wireless sensor networks: A survey. *IEEE Communications Surveys & Tutorials* 12, 2 (2010), 159–170.