Semantic Web 1 (2016) 1–5
IOS Press

Semantic Prediction Assistant Approach applied to Energy Efficiency in Tertiary Buildings

Iker Esnaola-Gonzalez [a,b], Jesús Bermúdez [b], Izaskun Fernandez [a], and Aitor Arnaiz [a]
[a] *IK4-TEKNIKER, Iñaki Goenaga 5, 20600 Eibar, Spain*
E-mail:{iker.esnaola, izaskun.fernandez, aitor.arnaiz}@tekniker.es
[b] *University of the Basque Country (UPV/EHU), Paseo Manuel Lardizabal 1, 20018 Donostia-San Sebastián, Spain*
E-mail: jesus.bermudez@ehu.eus

Abstract. Fulfilling occupants' comfort whilst reducing energy consumption is still an unsolved problem in most of tertiary buildings. However, the expansion of the Internet of Things (IoT) and Knowledge Discovery in Databases (KDD) techniques lead to research this matter. In this paper the EEPSA (Energy Efficiency Prediction Semantic Assistant) process is presented, which takes leverage of the Semantic Web Technologies (SWT) to enhance the KDD process for achieving energy efficiency in tertiary buildings while maintaining comfort levels. This process guides the data analyst through the different KDD phases in a semi-automatic manner and supports prescriptive HVAC system activation strategies. That is, temperature of a space is predicted simulating the activation of HVAC systems at different time and intensities, so that the facility manager can choose the strategy that best fits both the user's comfort needs and energy efficiency. The proposed solution is abstract enough to reuse it in similar use-cases of the same domain and it has been proved that improves the accuracy of predictions.

Keywords: Semantic Web Technologies, Knowledge Discovery in Databases, Energy Efficiency, Buildings

1. Introduction

Concerns over changing climatic conditions (i.e. global warming, depletion of ozone layer, etc.), energy security, and adverse environmental effects are growing among governments, researchers, policy makers, and scientists in developed as well as developing countries [61]. In order to meet the energy sustainability and minimize the climate change, the European Commission agreed a set of binding legislation inside the EU 2020 package. One of the spotlighted sectors regarding this package is the building sector which, according to the UNEP (United Nations Environment Programme) consumes about 40% of global energy and is responsible for the 36% CO2 emissions in the EU. Therefore, efficient management of building energy plays a vital role and is becoming the trend for future generation of buildings.

However, energy efficiency is not the only concern related with buildings. Since approximately 90% of people spend most of their time in buildings, indoor comfort is a must and poses a huge impact to preserve inhabitant's health, morale, working efficiency, productivity and satisfaction. As a consequence, it is necessary a system which fulfils the occupants' expected comfort index whilst reducing energy consumption during the operation of building.

In this context, the expansion of the Internet of Things (IoT) and Knowledge Discovery in Databases (KDD) techniques will lead to both researching the reduction of such prominent impact and the improvement of comfort levels. The KDD can be understood as a five steps process leading the extraction of useful knowledge from raw data [22], applicable for in-

stance in decision support systems. The five steps can be summarized as it follows:

1. Selection of datasets and subset of variables or data samples on which discovery will be performed.
2. Preprocessing tasks to ensure data quality and preparation for a subsequent analysis.
3. Transformation or production of a projection of the data to a form that data mining algorithms can work and improve their performance.
4. Data mining by selecting the algorithm that best matches the user's goals and their application to search for hidden patterns.
5. Interpretation and evaluation of the results, patterns and models derived, and application of them to make necessary decisions.

This process can involve significant iteration and can contain loops between any two of the mentioned steps as can be seen in Figure1.

In this paper the EEPSA (Energy Efficiency Prediction Semantic Assistant) process is presented. EEPSA process takes leverage of the Semantic Web Technologies (SWT) to enhance the KDD process for achieving energy efficiency in tertiary buildings. For that purpose expert knowledge in buildings, deployed devices and observations are used. The proposed process assists the data analyst during the different KDD phases, the robustness and performance of machine learning algorithms applied in the data mining phase are improved, and it eases the interpretation of the obtained results.

The rest of this paper is structured as follows. Section 2 introduces the related works and analyses existing main ontologies in the field. The EEPSA process is presented in section 3 and section 4 shows the application of this process in a real-world use case. Obtained results are evaluated in section 5 and finally the conclusions of this work are shown in section 6.

2. Related Works

2.1. KDD for Energy Efficiency in Buildings

KDD have traditionally been used to achieve energy efficiency in buildings such as in [27], where Artificial Neural Networks (ANN) and historic values have been used for short-time load forecasting in buildings. However, existing BMS (Building Management Systems) have generally failed to fully optimize energy consumption in buildings. [29] states that current and future information about events and weather (e.g. rain or snow) would help increasing the stability of the control systems minimizing energy consumption and increasing the occupants comfort. External meteorological conditions are used to improve the energy usage predictions in [62]. But it has been proved that not all external weather factors have the same impact in the energy consumption forecasting in buildings. In the use case presented in [44], effects of humidity and solar radiation have resulted to be less significant than external temperature.

Related work in [42], [58] and [68] shows that not only external climatologic factors affect the energy use in buildings. Most modern buildings still condition rooms assuming maximum occupancy rather than actual usage. As a result, rooms are often over conditioned. [20] proposes different HVAC control strategies based on occupancy prediction of rooms. In a similar way [57] focuses on a better heating scheduling by predicting future occupancy. Wireless motion sensors and door sensors are used in [40] to infer occupants presence and activate or deactivate HVAC systems accordingly. [48] aims at developing predictive control strategies that use both weather and occupancy forecasts to limit peak electricity demand while maintaining high user comfort.

So far, it has been proved that meteorological factors as well as occupancy of buildings have a significant impact both on the building energy consumption and comfort. The HVAC control strategies have also been deeply studied as a measure to achieve these two goals. However, the process of combining all these data sources into the KDD for exploiting them poses a big challenge. This research proposes the use of SWT towards the improvement of the whole KDD process and obtained results.

2.2. Semantic Web Technologies for KDD

In the last years, advantages of semantic technologies for data understanding as well as for the data mining process itself have been highlighted in [35] and [52]. Furthermore, many approaches have proposed the use of Semantic Web data to enhance different KDD phases. Semantic Web Technologies address how one would discover the required data in today's chaotic information universe, how one would understand which datasets can be meaningfully integrated, and how to communicate the results to humans and machines alike. That is why making sense of data and

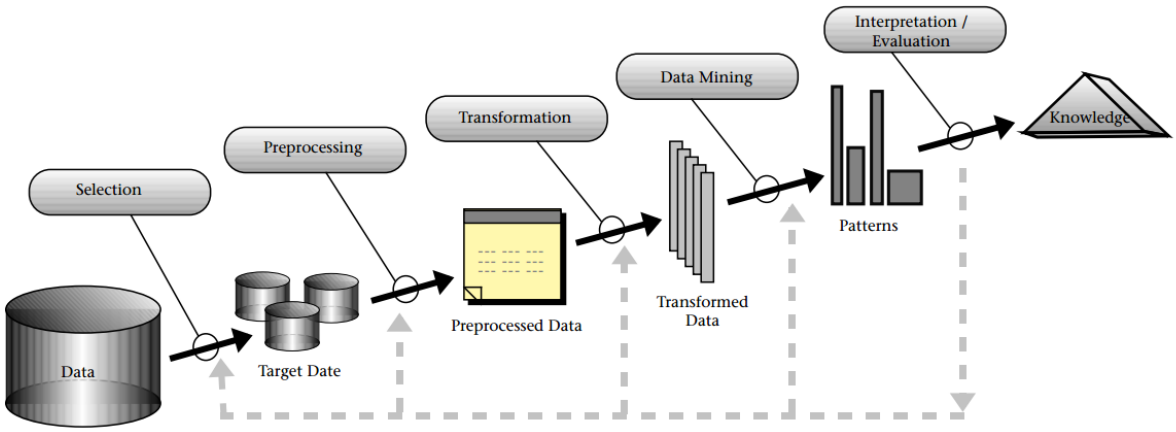

Fig. 1. An overview of the steps that compose the KDD Process [22].

gaining new insights works best if inductive and deductive techniques go hand-in-hand.

[17] states that the Internet of Things (IoT) and Open Data are particularly promising in real time predictive data analytics for effective decision support, and declares that the dynamic selection of Open Data and IoT sources for that purpose is the main challenge. Data quality is tackled in [23], [24] and [25], where data quality problems in Semantic Web data are identified by means of data validation rules. A review of the existing data quality works based on ontologies for the health domain is shown in [39]. In [53] desiderata and challenges for developing a framework for unsupervised generation of data mining features from Linked Data are identified. [43], [49] and [56] are examples of systems for enriching data with features that are derived from LOD. In [66] a feature-selection method based on ontology is proposed. The data mining environment RapidMiner [33] includes a Linked Open Data extension which provides a set of operators for augmenting existing datasets with additional attributes from open data sources [54]. In [46] semantic technologies are used to assist data scientists in selecting appropriate modelling techniques in the field of Statistics or Machine Learning and building specific models as well as the rationale for the techniques and models selected. [31] presents an ontology to support the meta-learning for algorithm selection in the data mining, while in [4] one of the first Intelligent Discovery Assistants is proposed. An overview of existing intelligent assistants for data analysis is provided in [59]. In [5] it has been noted that SWT can also have a potential impact in the Decision Support.

A detailed and extended survey on SWT within the KDD process can be found in [55]. The survey shows that, while many impressive results can be achieved already today, the full potential of Semantic Web Technologies for KDD is still to be unlocked.

This research contributes at exploiting SWT to enable an improved KDD process for energy efficiency in tertiary buildings. The proposed EEPSA process guides the data analysts through the KDD phases and enables its automation. Besides, it enriches data, improves its quality and generates new features, which results in more robust solutions. The use of SWT also aids at the achievement of a higher abstraction level, enabling the reuse in similar use cases.

2.3. Existing Ontologies in the Field

BIM (Building Information Modelling) deals with the representation of a functional and physical characteristics of a building [18]. That is, in a BIM model static information of a building element can be queried, such as a door: its material, when it was installed, or even the changes the door received until date. But for instance, it is not possible to know whether the door is opened or closed in a given moment. That is why, in order to transform the building static data into live data, it is necessary to integrate information coming

from IoT and sensing device network nodes. This data integration across several data sources can be obtained by adopting SWT, namely ontologies.

Keeping this in mind, a wide revision of ontologies in the field has been performed. Below, a brief summary of the most relevant ontologies of the current research domain is presented. Other ontologies such as Semanco [41] or the Aemet Network of Ontologies [3] have also been analysed, but are not covered in such a depth. Some of the consulted surveys to identify these ontologies have been [19] and [36]. An interesting comparison between different IoT ontologies is also covered in [60]. The catalogues Linked Open Vocabularies [65] and LOV4IoT [28] have been used to search vocabularies covering desired concepts.

2.3.1. ifcOWL Ontology

IfcOWL provides an OWL representation of the Industry Foundation Classes (IFC) Schema which is the open standard for representing building and construction data. Using the ifcOWL ontology, one can represent building data in directed labelled graphs (RDF) [50]. The graph model and the underlying web technology stack allows building data to be easily linked to material data, GIS (geographic information systems) data, product manufacturer data, sensor data, classification schemas, social data and so forth.

The ifcOWL ontology aims at supporting the conversion of IFC instance files into equivalent RDF files. That is, it is of secondary importance that an instance RDF file can be modelled from scratch using the ifcOWL ontology and an ontology editor. Since modelling a space from scratch is not a straightforward task, further modifications and extensions have been proposed in order to tackle this situation [51]. For example [16] presents ifcWoD (IFC Web of Data) ontology, which has several advantages compared with ifcOWL ontology: it simplifies and eases query writing, query response time for retrieving building data is improved and decreases data redundancy. However, this ontology is not available at the moment of writing this article.

2.3.2. DogOnt Ontology

It allows to formalize all the aspects of IDEs (Intelligent Domotic Environment) and it is designed with a particular focus on interoperation between domotic systems [6]. Mainly covering device, state and functionality modelling, it also supports device independent description of houses, including both controllable and architectural elements. DogOnt provides different reasoning mechanisms corresponding to different goals: to ease the model instantiation (by means of a set of auto completion rules), to verify the consistency of model instantiations, and to automatically recognize device classes starting from device functional descriptions.

However, other information such as measurements or insulation of building elements is not described in DogOnt. Observations made by sensing devices which are essential for a KDD process in the energy efficiency context, are not covered either.

2.3.3. SSN Ontology

The Semantic Sensor Network (SSN) ontology is developed by the W3C Semantic Sensor Networks Incubator Group (SSN-XG) and can describe sensors, accuracy and capabilities of such sensors, observations and methods used for sensing [11]. Also concepts for operating and survival ranges are included, as these are often part of a given specification of a sensor, along with its performance within those ranges. Finally, a structure for field deployment is included to describe deployment lifetime and sensing purpose of the deployed instruments. The SSN ontology is aligned with DOLCE ultra-lite (DUL) ontology and built around a central Ontology Design Pattern (ODP) describing the relationships between sensors, stimulus, and observations called the Stimulus-Sensor-Observation (SSO) pattern.

The SSN ontology does not contain properties which can be measured by sensors. Neither is covered related material such as units of measurements of these properties, locations or hierarchies of sensor types, and time-related concepts. All these knowledge has to be modelled or imported from other existing vocabularies.

2.3.4. SAREF Ontology

The Smart Appliances REFerence (SAREF) ontology is a shared model of consensus that facilitates the matching of existing assets in the smart appliances domain [14]. The ontology is based on the fundamental principles of reuse and alignment of concepts and it also provides building blocks that allow separation and recombination of different parts of the ontology depending on specific needs.

SAREF enables modelling devices and sensors in terms of functions, states and services they provide. Nevertheless, the ontology does not address the problem to describe the observation in an interoperable manner to ease further tasks such as reasoning. It provides the link to FIEMSER data model covering building-related concepts but this knowledge is not

enough to describe building elements and their features.

SAREF4BLDG ontology[1] presents an extension of SAREF for the building domain based on the IFC standard. It is limited to the description of devices and appliances within the building domain, so building elements and their features are not covered. However new classes such as buildings, spaces and the physical objects they contain are described.

2.3.5. FIESTA-IoT

FIESTA-IoT Ontology aims to achieve semantic interoperability among heterogeneous test beds [2]. The ontology takes inspiration from the methodologies for reusing and interconnecting existing ontologies. To build the ontology, ontologies and taxonomies, such as Semantic Sensor Network (SSN), M3-lite (a lite version of M3 ontology), WGS84, IoT-lite, OWL-Time, and DUL ontology have been reused.

Despite sensing devices are deeply described and covered, tagging and actuating devices are not at the same level. Besides, even though the smart building domain is described, building elements and its features are not.

2.3.6. IoT-O Ontology

It is a core-domain modular IoT ontology proposing a vocabulary to describe connected devices and their relation with their environment [60]. It is intended to model knowledge about IoT systems and to be extended with application specific knowledge. It has been designed in separated modules to make the reuse and/or extension of it easier. It is constituted of five different modules:

- A sensing module, based on SSN Ontology
- An acting module, based on SAN (Semantic Actuator Network)
- A service module, based on MSM (Minimal Service Model)
- A lifecycle module, based on a lifecycle vocabulary and an IoT-specific extension
- An energy module, based on PowerOnt [7]

The building information is described reusing DogOnt concepts and information regarding building elements or their features is not covered.

2.3.7. SmartHomeWeather Ontology

Smart Home Weather is an OWL ontology that covers both the weather data and the concepts required to perform weather-related tasks within smart homes [63]. Apart from concepts such as weather phenomena and states that can be used to model external climatic condition, the ontology covers weather forecasting data over a time range that it is suitable to use within a smart home.

3. EEPSA in KDD Support

Nowadays data analysts receive no guidance in the KDD processes and consequently, novice analysts are typically completely overwhelmed. They have no idea which variables and tasks can be confidently used, and often resort to trial and error. Besides, being a non-expert in the domain complicates even more the process to make accurate predictions. Therefore, there is a dire need to support both data analyst expert and novices during the whole KDD process.

The EEPSA process makes use of SWT as a contribution to overcoming this hurdle in the domain of energy efficiency in tertiary buildings. First of all, data is linked to the EEPSA ontology[2]. This ontology aims to capture all the necessary expert knowledge for the EEPSA process mainly related to buildings, sensing and actuating devices, and their corresponding observations and actuations. This linking phase is fundamental for enriching data, integrating heterogeneous data and representing it in a more domain-oriented way, as well as for enabling the improvement of the upcoming KDD phases. In the data selection phase the data analyst is assisted to decide which might be the most relevant variables for the matter at hand. The preprocessing phase takes leverage of a predefined set of SPARQL rules to detect outliers and propose possible methods to solve them according to their potential cause. The transformation phase generates additional knowledge in form of new attributes. All these tasks contribute to improve the robustness and performance of machine learning algorithms applied in the data mining phase and it eases the interpretation of the obtained results. Besides, the proposed process is reusable in similar use cases of the same domain due to its high abstraction level.

The EEPSA process tackles the achievement of energy efficiency while maintaining users' comfort in ter-

[1] https://w3id.org/def/saref4bldg

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

tiary buildings. There are many complementary ways to save and optimize energy use in buildings, but since temperature is the most important weather parameter affecting electric load, forecasted temperatures constitute a basic ingredient in energy efficiency plans [1]. However, it is important to make clear that temperature forecasting is not the goal of the EEPSA process. These predictions are used to support prescriptive HVAC system activation strategies. That is, temperature of a space is predicted simulating the activation of HVAC systems at different time and intensities. For example, prediction of the temperature in a room when all HVAC systems are activated four hours in advance, when half of existing HVAC units are activated six hours in advance, etc. Estimating the temperature obtained with different strategies in advance, the one that uses energy in a more efficient way while maintaining the optimal comfort[3] can be chosen.

The EEPSA process targets different KDD phases.

3.1. Linking

This previous phase to the KDD process which can also be transversal, consists in annotating data semantically to enrich it for the following phases. When linking or mapping raw data to existing ontologies or vocabularies a better representation of data is achieved, structuring it and setting formal types and relations between concepts. Data integration is also achieved [45], and additional background knowledge can be added to the dataset. Besides, the resulting dataset avoids semantic interoperability issues [47], providing both human and machines with the same meaning of terms. This increases the dataset value and the potential to improve the upcoming KDD phases.

In the energy efficiency in buildings domain, there are three main information sources to be annotated: (i) the space in which the energy efficiency is going to be performed, (ii) the devices deployed in it, and (iii) the information gathered by those devices. Data coming from these sources has to be unified and EEPSA ontology plays a key role for that purpose.

The ontology DogOnt has not been considered appropriate to reuse building-related terms, because it targets residential buildings (which differ from tertiary buildings) and building element's coverage is not as broad as needed. IfcOWL represents the IFC open standard for building and construction data, but it does not enable a straightforward modelling of a building from scratch. Since the data-analysts using the EEPSA process might not be experts in the building domain, it has been decided to offer a simpler way to model a space. That is why core concepts to model a building's space have been created such as building elements (e.g. walls, doors or windows) and their features (e.g. area, thickness). A set of IFC property sets (e.g. *pset_DoorCommon isExternal*) have been adapted to ease their modelling, based on [16] proposal. In future stages of the research, it is expected to translate all the building modelling concepts defined in the EEPSA ontology into the ifcOWL representation, in order to enable interoperability with other processes.

Sensors and their measured observations have been described reusing terms of the SSN Ontology[4] such as *ssn:Sensor* and *ssn:Observation*, as well as the SSO pattern. For actuators and their actuations' descriptions, the Actuation-Actuator-Effect pattern defined in the SAN (Semantic Actuator Network) ontology[5] and classes like *san:Actuator* and *san:ActuatingDevice* have been reused. Properties (e.g. *m3-lite:Temperature*), units of measurements (e.g. *m3-lite:DegreeCelsius*) and sensing devices (e.g. *m3-lite:Thermometer*) have been annotated with a module extracted from the M3-lite ontology, with the tool Locality Module Extractor[6] [12]. Other widely used ontologies have also been reused to represent spatially located things (Basic Geo Vocabulary[7] with concepts like *wgs84_pos:point* and *wgs84_pos:location*) and time-related entities (OWL-Time Ontology[8] and resources like *time:Instant* class and *time:hasDateTime* property).

Besides the aforementioned integration and interoperability advantages, the resulting data is more domain oriented than the original source, and makes the solution more application-independent. Consequently, after the annotation there is no need for the data analyst to be aware of the structure of the underlying raw data.

Whether the annotated data is stored natively as RDF or viewed as RDF via middleware, SPARQL queries will be later used to access data across diverse data sources.

[3] The optimal comfort can be understood in many ways: a temperature that ranges between some given values, a temperature that suffers less variation during a period of time, etc.

[4] https://www.w3.org/TR/vocab-ssn/

[5] https://www.irit.fr/recherches/MELODI/ontologies/SAN.owl

[6] https://www.cs.ox.ac.uk/isg/tools/ModuleExtractor/

[7] http://www.w3.org/2003/01/geo/wgs84_pos

[8] http://www.w3.org/2006/time

Summarizing, after the data analyst links data to the EEPSA ontology, data integration, interoperability and independence from original source are improved.

3.2. Data Selection

This is the first phase of a KDD process where relevant datasets and subset of variables are selected to eventually perform knowledge extraction over them. In order to do that, the data analyst has to understand what knowledge is captured in the data and which the additional knowledge that can be extracted from the data is. However, this step is often not trivial and in most cases, a domain-specific knowledge is needed to successfully complete it.

Existing works focus on the use of tools and approaches to visualize and explore LOD to understand data [13]. However, no relevant work that supports the data analyst in data selection phase has been spotted. In the EEPSA process, SWT are used to support the data analyst choosing the most relevant datasets and variables related with the energy efficiency problem at hand.

Apart from core concepts to model buildings and their spaces, EEPSA ontology also describe different space-types, characterized by features like insulation, location and building elements. For example, a space located in an underground floor of a building belongs to class *eepsa:BelowGroundLevelSpace*.

Each type of space has assigned a set of variables that might affect its indoor conditions with the property *eepsa:isAffectedBy*. For example, a space with windows towards the outside, is a naturally enlightened space (*eepsa:NaturallyEnlightenedSpace*) and it is inferred that, apart from its indoor factors (such as temperature and humidity) might have its indoor temperature affected by the received solar radiation, as well as the elevation and direction of the sun. Consequently, although not being an expert in the domain, thanks to the OWL definitions within the EEPSA ontology the data analyst will get to know which type of space is the one being analysed and the variables affecting it.

The process of selecting a data source is subjective based on the needs of the consumer [17]. That is why although a subset of variables is suggested, the data analyst is the one who must decide whether to choose them or other ones.

Summing up, once the data analyst has the target space semantically annotated (Linking phase) and thanks to the knowledge captured in the ontology and reasoning capabilities, the data selection phase classifies the annotated space into one or various space-types. Because the target space is classified as a specific space-type, it is inferred that it might be affected by some specific variables. Consequently, the data analyst will get to know which variables might be relevant for the target space, even though not being an expert in the domain. Besides, this task is semi-automatized.

After having suggested which variables are the most relevant ones for the task at hand, the data analyst needs to know which of them are being collected by the devices or other mechanisms deployed on the space and which are not. This can be obtained with a SPARQL query (see Appendix A, Listing 1) and thanks to the previous Linking phase, where all data has been semantically annotated.

EEPSA process uses OWL inferences to assist the data analyst in classifying the space at hand and suggesting variables affecting it. Besides, SPARQL queries is also provided in order to know whether those variables are being collected by devices or not.

The next step deals with preprocessing collected data in order to ensure their quality.

3.3. Preprocessing

Today's real-world datasets are highly susceptible to noisy, missing, and inconsistent data due to their typically big size and their likely origin from multiple, heterogeneous sources [30]. Low-quality data will lead to low-quality mining results, that is why it is important to ensure data quality in KDD processes. There are several data preprocessing techniques to increase data quality, which can consequently improve the accuracy and efficiency of data mining algorithms. Besides, these techniques are not mutually exclusive and may be applied together.

3.3.1. Outlier Detection

Outliers are data objects that stand out amongst other data objects and do not conform to the expected behaviour in a dataset [37]. In addition, outliers can complicate the knowledge extraction process and lead to wrong conclusions.

Outlier detection is the process of finding anomalies in a dataset and is an essential task in a wide range of domains including fault detection in safety critical systems, intrusion detection for cyber-security and fraud detection for credit cards. This process has been a widely researched topic for many years and there has been an abundance of work from statistics, ge-

ometry, machine learning, database, and data mining communities. There are many outlier detection methods divided into groups according to their assumptions regarding normal data objects versus outliers such as model-based, distance-based or density-based. Further information regarding these and other outlier detection methods can be found in [9] and [32].

However, most of these conventional methods might not be directly applied to outlier detection tasks in scenarios where context is determinant to decide whether a data object is an outlier or not. For example, scenarios where an observation may be considered an outlier in one context (e.g. 40°C is an outlier for a winter day in the north of Spain), but not an outlier in another context (e.g. 40°C is not an outlier for a summer day in the south of Spain). Besides, identifying the potential cause of outliers still remains an unsolved challenge in most cases, even though it could be very helpful for determining how to act on the detected outliers.

Although being a very researched topic, outlier detection has not received sufficient attention from the Semantic Web Community. One of the scarce solutions where these technologies have been employed is [67], where a domain ontology has been used as a support to apply a conventional outlier detection method. The SWT could have a prominent impact to contribute improving both the detection and classification of outliers.

With regards to wireless sensor networks (WSN), which are an essential component to capture building conditions, several factors make them prone to outliers due to their particular requirements, dynamic nature and resource limitations [21]. Apart from these factors, WSNs are also context dependent, so that results obtained after applying conventional techniques might be skewed. That is why the EEPSA process proposes a novel outlier detection method based on knowledge.

Thanks to the linking phase, it is possible to get additional information about an observation such as the sensing device that observed it or its location. Supported by this additional information, it is possible to establish the context in which the observation has been made and exploit this knowledge. OWL descriptions may not suffice to produce all desired inferences of some of these contexts, so an alternative paradigm for knowledge modelling is needed. SPARQL 1.1 Update, an update language for RDF graphs has been used to model expert knowledge and to define a series of rules in order to determine if a data object is an outlier or not according to the context in which takes place. Rules have associated an expression in natural language so that the data analyst can understand them and choose a subset of these rules to execute.

The class *eepsa:Outlier* is defined as a subclass of *ssn:Observation* in the EEPSA ontology in order to represent observations that do not conform to the expected behaviour. In addition, these SPARQL rules are designed to classify the detected outliers according to their potential cause. That is why a hierarchy of outlier types is defined, covering contextual outliers such as the ones caused by the bad location of a sensing device (*eepsa:OutlierCausedByDeviceLocation*) or by an error of the device itself (*eepsa:OutlierCausedByDeviceError*). For example, imagine a temperature sensor located in the open air where it is exposed to the rain. Wet sensors' temperature and humidity measurements can be very different due to the evaporation of water from its surface. This expert knowledge is used to define a SPARQL rule, namely the rule "Detection of anomalous observations caused by rain" (see Appendix A, Listing 2). After executing this rule, temperature or humidity observations measured by that sensor that fulfil the SPARQL rule's conditions will be considered outliers and classified as *eepsa:OutlierCausedByRain*.

Outliers can occur for various reasons and understanding them might help determining what action to perform. That is why each outlier class has assigned a proposed method to offset the problem with the property *eepsa:hasSolvingMethod*. For example, a temperature outlier caused because the sensing device got heated by direct sunlight (*eepsa:OutlierCausedBySunlight*) has assigned two recommended solution methods: *eepsa:DeviceRelocation*, which recommends to relocate the device to an adequate place where it is not hit by sun and *eepsa:DeviceShelter*, recommending to shield the device with a Stevenson Screen or a similar one to cover it from direct heat radiation. Following any of these advices should avoid the device getting heated by direct sunlight and measuring erroneous observations.

In this stage, EEPSA process provides the data analyst a set of SPARQL rules to detect outliers within a dataset and classify them according to their potential cause. OWL inferences are also used to propose methods to solve the cause of the outliers and avoid measuring them in the future.

Once the existing data is preprocessed and the data quality has been ensured, the next step in the KDD process is the Transformation phase.

3.4. Transformation

In this stage, a projection of the data is produced into a form in which data mining algorithms can work. Amongst all the possible tasks in the Transformation phase, the EEPSA process focuses in the feature generation task. Other tasks such as feature extraction are expected to be studied in future stages of the research.

The vast majority of existing feature generation solutions such as [10], [49] and [43] choose a general knowledge base like DBpedia and YAGO to obtain properties about the mapped entities and generate new attributes. This approach is considered to only exploit SWT capabilities partially, so other alternatives are proposed: the generation of new features from domain-specific knowledge bases and the inference of new features based on existing data.

Variables that are observable by sensors and in which actuators have an effect, are described within the EEPSA ontology. For each of them, the sources they can be retrieved from are captured with the *eepsa:hasDataSource* property. This can be useful for cases when there is a concrete property that is not being collected in the target space. For example, a badly insulated space (*eepsa:BadlyInsulatedSpace*) might be affected by outdoor humidity among others. If there is no sensing device observing this variable, data values can be retrieved from a nearby weather station. So, when looking for a specific variable which is not being monitored, the captured knowledge lets the data analyst know which the data sources of that variable can be.

Nowadays, with the advent of (Linked) Open Data repositories, data can be retrieved from many trustworthy third-party RDF Stores. In the building scenario, where it has been proved that external meteorology affects the energy consumption, weather services enable the possibility of increasing datasets value with specific knowledge.

However, there are variables that cannot be obtained from third party data sources. For those cases, an alternative is offered in the form of rules. For example, indoor illuminance values in spaces with windows next to the outside (*eepsa:NaturallyEnlightenedSpace*) can be derived from the sky's cloud cover, sun elevation and direction information.

The proposed feature generation phase has to be performed at least as many times as the number of variables to generate. The goal is to get the variables previously suggested in the Data Selection phase for the upcoming data mining phase. Retrieved or inferred data is considered to have a minimum quality, so preprocessing tasks should not be necessary afterwards.

EEPSA process uses OWL inferences to identify sources of information where certain variables can be retrieved from. Besides, data analyst also can use SPARQL rules to infer values of variables based on existing data.

3.5. Data Mining

This is the phase where intelligent methods such as machine learning algorithms are applied to extract knowledge. The data analyst will try to make the best predictions with views to achieving energy efficiency in the target space. For that purpose, data enhanced in previous phases has to be retrieved and integrated in the data analysis environment, mainly by means of SPARQL queries.

3.6. Interpretation

Interpreting results obtained from the data mining phase is not always a straightforward task. Many times, even being an expert in the domain is not enough to understand the results. If underlying semantics of data are not correctly interpreted, results may not be as precise and consistent as they can be [38].

In [15] and [64] Linked Open Data has been proposed as a source of additional information to support the interpretation of the results for the data mining method. However, an effective decision-making must result from reasoning and analysis of knowledge, and must also take into account the experience and expertise of decision-makers. The EEPSA ontology is intended to be extended with this knowledge in further stages of the research, in order to contribute in the Interpretation phase. Besides, thanks to the linking phase, data is enriched so that additional information about the domain can be brought, which contributes to an easier and more effective results interpretation.

4. EEPSA on the Loop

The feasibility of the EEPSA process has been tested in the IK4-TEKNIKER building, a technological centre constituted as a not-for-profit foundation located in Eibar (Basque Country, Spain). The scenario in which the EEPSA process has been applied is the second floor of this building (from now on referred as Open Space) which is a single big room without walls

that acts as an office and over 200 people work on a daily basis. The Open Space is equipped with sensing devices developed in Tibucon project[9] that observe temperature, humidity and illuminance with a periodicity of 5 minutes. Information regarding the deployed HVAC systems is also collected.

A service is needed for suggesting the facility manager when HVAC systems have to be activated in order to reach a minimum temperature of 23°C at 08:00 a.m. (the time when workday starts). HVAC activation strategy needs to be efficient from an energy expense point of view too. The EEPSA process has been applied to meet facility manager's requirements.

The first step of the process is the linking phase. All data regarding deployed devices and their gathered observations are stored in a PostgreSQL Database. In order to semantically annotate this data to the EEPSA ontology, Ontop tool[10] has been used. Ontop is an OBDA (Ontology-Based Data Access) tool which enables mappings between relational DB and an ontology [8]. It also enables to build a semantic layer, so that data can be queried with SPARQL language while staying as relational DB. Mappings can be implemented using the Ontop Protégé plugin. Nevertheless, since further inference capabilities than the ones offered by Ontop are needed, it has been decided to dump RDF assertions derived from mappings. These assertions have been stored in a Virtuoso server 07.20.3217 version, running on an Ubuntu 14.04 Server. Since it is considered to contain sensitive data, this RDF Store is private. The Open Space has also been modelled using the EEPSA ontology.

Once the Open Space itself, the deployed devices and their observations have been semantically annotated, it is time for the data selection. In order to make predictions as accurate as possible, variables that affect energy consumption of the Open Space have to be identified. But before that, it has to be determined what type of space is the Open Space. According to what is inferred from EEPSA ontology class definitions, the Open Space is an adjacent to the outside (*eepsa:AdjacentToOutsideSpace*) and naturally enlightened (*eepsa:NaturallyEnlightenedSpace*) space. As a result of the definition of these space type classes, it is inferred that indoor temperature might be affected by variables such as occupancy, indoor and outdoor temperature and solar radiation. However, after executing a previously defined SPARQL query that checks the observed properties in a space, it is concluded that not all of these are being observed.

The next step is to ensure the quality of the collected observations. These observations are queried accessing the private RDF Store where they have been previously stored. During preprocessing phase, a subset of SPARQL rules provided by the EEPSA process has to be selected in order to detect outliers among the gathered observations. Afterwards, additional information can be queried, such as the number of detected outliers, the potential cause for being an outlier, and the possible solution. Results obtained after applying these rules are in section 5.3.

Once the data preprocessing has finished it is time to get those variables affecting energy consumption of the Open Space but that are not currently being measured. It is necessary to see how to obtain them, in order to enable the feature generation. That information is also inferred thanks to the knowledge captured in the EEPSA ontology.

Weather stations regulated by Euskalmet (Basque Meteorology Agency) and the observations they measure have been made publicly available[11] in a Virtuoso Open Source version 07.20.3217 Server[12]. The data analyst can first of all use a predefined GeoSPARQL query to identify the closest weather stations gathering the desired variables. The query is parameterizable, so that the analyst just needs to put the coordinates of the building and select the wished variable(s). The query returns a set of weather stations, sorted by proximity to the target building. However, it is not compulsory for the data analyst to choose the closest weather station. Other factors rather than the distance can influence on the election of one or another weather station, such as the altitude where the sensing device is deployed. These information is also collected and can be queried. Once the data analyst has decided which weather station chooses to retrieve the data from, a parameterizable SPARQL query has to be performed over the same endpoint. This time, the data analyst needs to determine the weather station, the variables and the time span to retrieve the needed information.

It has been determined in the data selection phase that one of the variables that affect the Open Space is the outdoor temperature. But due to its low quality, it

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

[10] http://ontop.inf.unibz.it/

[11] All data has been provided by Open Data Euskadi (Basque Open Data portal) and Euskalmet, semantically annotated with the EEPSA ontology, and published in the RDF Store.

[12] http://193.144.237.227:8890/sparql

has been decided to retrieve this information from a higher quality data source. It is inferred that the outdoor temperature can be obtained from a weather station, so first step is to check if there are any weather stations nearby. That information is contained in the same SPARQL endpoint and is retrieved executing a GeoSPARL query parametrized with Open Space location and the desired variable (see Appendix A, Listing 3). This query returns the closest weather stations to IK4-TEKNIKER building measuring the outdoor temperature (see Figure2).

The weather station named Eitzaga is chosen to retrieve the outdoor temperature information[13]. A SPARQL query parametering the variable (outdoor temperature), weather station (Eitzaga) and the timespan (between 1st January and 1st of August of 2016) is executed for that purpose.

After repeating this feature generation task as many times as needed, all data has to be used in the following data mining phase. In this case, the RapidMiner Studio 7.1 version have been used alongside with the Linked Open Data extension. Within this extension, the operator SPARQL Data Importer has been used to query the RDF Store and retrieve the information. It has also been used the Series extension in order to work with time series. After testing different algorithms, the Vector Linear Regression with its default parameters has been chosen for building the predictive model.

5. Evaluation and Results Discussion

5.1. Experimental Setup

Available data in the Open Space, comes from three Tibucon sensing devices located indoors, a Tibucon sensing device located outdoors, and HVAC systems. Tibucon devices measure temperature, humidity and illuminance, and HVAC systems information has been simplified to whether any HVAC unit is activated or not.

A baseline experiment was executed without the use of EEPSA process. This baseline's results were compared with the ones obtained after applying the EEPSA process, to see if they were improved and to what extent. Collected data ranged between 31st January 2016 and 1st August 2016 (6 months), and it was sampled hourly. Around 20% of data was not measured due to external problems and in many circumstances there were high temperature values which were unlikely. So these data quality problems had to be handled. Based on a previously performed set of experiments, the Vector Linear Regression algorithm was used to create the baseline's predictive model. It was also proved that best results are obtained using a smaller window size, which can be caused by the small quantity of available data (6 months).

[13] A study of nearby weather stations has been conducted and it concluded that this one is the most suitable one.

5.2. Evaluation

Performance of the forecaster is characterized by two statistical estimates: the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Measures based on percentage errors (e.g. Mean Absolute Percentage Error, MAPE) was dismissed because of their disadvantage of being infinite or undefined if data is zero, and having extreme values when close to zero. Therefore, a percentage error makes no sense when measuring the accuracy of temperature forecasts on the Fahrenheit or Celsius scales[34].

Baseline experiment was developed without the support of the EEPSA process. Some predictive models were built using different combinations of available variables and fine-tuning the parameters for their window-sizes. Best results were obtained with a model containing a window of 553 features: a window of past 504 indoor temperature observations, 24 past external temperature, a feature representing the HVAC value, and another one for the date time. Predicted temperatures for the future 24 hours had a MAE of 0.70°C and a RMSE of 0.87°C.

For the EEPSA-enabled model, first of all the Linking phase was applied. Then, EEPSA data selection suggestions were taken into account and the outlier detection task was applied in observations gathered by devices. Thanks to the generation of new attributes, available data pool became larger. Variable selection and their window-sizes was fine tuned to create a model that accurately predicts Open Space's upcoming 24 indoor temperatures. Most accurate model was built with 168 past indoor temperatures, 24 past observations for outdoor temperature, outdoor humidity, outdoor wind speed and HVAC status, 2 features to describe current space occupancy, and 4 features describing the date (month, hour, day of the week and date time). Temperature predictions obtained with this model had a MAE of 0.54°C and a RMSE of 0.67°C.

stationID	stationName	city	owner	distanceToBuilding
"C075"	"Eitzaga"	"Zaldibar"	"http://es.dbpedia.org/page/Euskalmet"	5.86976
"C0D3"	"Aixola (Embalse)"	"Elorrio"	"http://es.dbpedia.org/page/Euskalmet"	6.91178
"C078"	"Altzola (Deba)"	"Elgoibar"	"http://es.dbpedia.org/page/Euskalmet"	8.17392
"C0BE"	"Berriatua"	"Berriatua"	"http://es.dbpedia.org/page/Euskalmet"	13.2363

Fig. 2. Closest Euskalmet weather stations to IK4-TEKNIKER building measuring temperature

5.3. Results Discussion

Obtained results improve the MAE in 0.16°C and RMSE by 0.20°C. Taking into consideration the small room for improvement of the baseline on the Open Space (where MAE is below 1°C), this is a considerable improvement of the prediction accuracy (over 20%). However, as stated along the article, the true impact of the EEPSA process should not be solely based on accuracy improvement.

Data Selection suggested incorporating to the predictive model variables that a data analyst not-expert in the domain may overlook, such as the temperatures registered outside the building.

Thanks to the knowledge-based outlier detection task, it was detected that the Tibucon device located outdoors, had 4,209 temperature observations that were anomalous. Out of them, 2,408 were of type *eepsa:OutlierCausedBySunlight* and the remaining 1,801 observations were outliers caused by rain. Although labelling all these data objects as outliers, they have been classified in different classes according to their potential provenance. This proves that the sensing device located outdoor gets hit by the sun in certain time spans and gets affected by precipitation in rainy days. Thanks to this information it has been decided to relocate the device in more adequate place where it is protected from direct sunlight and against rain.

The generation of new features enabled the improvement of predictions in certain days. The Open Space has split shift schedule from Monday to Thursday but there are special days like the 23rd March 2016 (Wednesday), where it was an intensive work day. With the application of EEPSA process, occupancy information is obtained and incorporated to the predictive model, which significantly reduces MAE of prediction for this day. Baseline model's MAE lowered from 0.64°C to 0.36°C and RMSE was also reduced from 0.84°C to 0.46°C when using the EEPSA supported predictive model.

Regarding the simulations, the models built without applying the EEPSA process showed that even in cold winter days, indoor temperature would increase from 05:00 a.m. on, even without activating HVAC system and without space occupancy. This simulation is incorrect and consequently facility manager's HVAC activation strategy cannot be based on it. Increasing the external temperature data quality and adding other factors contributed in a more accurate simulation of different scenarios.

6. Conclusions

6.1. Benefits of the EEPSA Process

The EEPSA process takes leverage of SWT to enhance the KDD process for achieving energy efficiency while maintaining comfort levels in tertiary buildings. Data analyst is guided through the different KDD phases in a semi-automatic manner, helping both novice and KDD experts. First of all, data is linked to the EEPSA ontology, which aims to capture all the necessary expert knowledge for the EEPSA process mainly related to buildings, sensing and actuating devices, and their corresponding observations and actuations. This linking phase is fundamental for enriching data, integrating heterogeneous data and representing it in a more domain-oriented way, as well as for enabling the improvement of the upcoming KDD phases. In the data selection phase the data analyst is assisted to decide which might be the most relevant variables for the matter at hand. The preprocessing phase takes leverage of a predefined set of rules to detect outliers and propose possible methods to solve them to ensure data quality. The transformation phase generates additional knowledge in form of new attributes. All these tasks contribute to improve the robustness and performance of machine learning algorithms applied in the data mining phase and it eases the interpretation of the obtained results. Besides, the proposed process is reusable in similar use cases of the same domain due to its high abstraction level.

6.2. Future work

The EEPSA process proposed in this paper contributes to raise awareness of the possibilities of the SWT. However, SWT can be further exploited to improve the EEPSA process, implementing some of the following tasks:

1. A rather simplified model is provided to the user for annotating the space where energy efficiency is going to be performed. To make the solution interoperable with other systems though, it should to be translated into the ifcOWL ontology. A tool fulfilling this objective should be developed.
2. EEPSA ontology should be completed with more detailed information to reflect the effect of features like materials or building envelope sealing. This is thought to enable a greater assistance during the KDD process.
3. Missing values in a dataset pose a problem for the data quality. After a study, it has been determined that necessary knowledge regarding missing values imputation methods [26] could be captured in order to suggest the most suitable one for each situation.
4. Interpretation phase has a big potential for exploiting semantics. Research on this topic should be conducted in further stages.

The EEPSA process is intended to be used by non-experts, so it should have an intuitive interface for every task. Currently the model of the target space has to be done manually and depending on the complexity of the space and the knowledge of the user, it can become a time-costing task. This task should be facilitated with a GUI where the user could add building elements and features to the space easily.

To test the reusability of the proposed EEPSA process, it is going to be applied in another tertiary building, namely in the Bilbao Exhibition Centre (BEC). The BEC is located in Baracaldo (Basque Country, Spain) and covers an area of 251.055 square meters distributed in six pavilions intended for exhibitions.

7. Acknowledgment

Part of the presented work is based on research contacted within the project BID3ABI (Big Data para RIS3 2016), which has received funding from the Basque Government (ELKARTEK 2016) under grant agreed reference KK-2016/00096. This work is also supported by FEDER/TIN2016-78011-C4-2-R.

We thank Euskalmet (Basque Meteorology Agency) for assistance with weather stations and observations, as well as Zuzenean (Basque Citizen's Advice Service) for helping us with Open Data Euskadi (Basque Open Data portal).

References

[1] R. Abdel-Aal, *Hourly temperature forecasting using abductive networks*, Engineering Applications of Artificial Intelligence 17 (2004) 543-556.

[2] R. Agarwal, D.G. Fernandez, T. Elsaleh, A. Gyrard, J. Lanza, L. Sanchez, N. Georgantas and V. Issarny, *Unified IoT Ontology to Enable Interoperability and Federation of Testbeds*, 3rd IEEE World Forum on Internet of Things (2016).

[3] G. Atemezing, O. Corcho, D. Garijo, J. Mora, M. Poveda-Villalón, P. Rozas, D. Vila-Suero and B. Villazón-Terrazas, *Transforming meteorological data into linked data*, Semantic Web 4 (2013) 285-290.

[4] A. Bernstein, F. Provost and S. Hill, *Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification, Knowledge and Data Engineering*, IEEE Transactions on 17 (2005) 503-518.

[5] E. Blomqvist, *The use of Semantic Web technologies for decision support - a survey*, Semantic Web 5 (2014) 177-201.

[6] D. Bonino and F. Corno, *Dogont-ontology modeling for intelligent domotic environments*, International Semantic Web Conference (2008) 790-803.

[7] D. Bonino, F. Corno and L. De Russis, *Poweront: An ontology-based approach for power consumption estimation in smart homes* , Internet of Things. User-Centric IoT, Springer, 2015, pp. 3-8.

[8] D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro and G. Xiao, *Ontop: answering SPARQL queries over relational databases*, Semantic Web (2016) 1-17.

[9] V. Chandola, A. Banerjee and V. Kumar, *Anomaly detection: A survey*, ACM computing surveys (CSUR) 41 (2009) 15.

[10] W. Cheng, G. Kasneci, T. Graepel, D. Stern and R. Herbrich, *Automated feature generation from structured knowledge*, Proceedings of the 20th ACM international conference on Information and knowledge management (2011) 1395-1404.

[11] M. Compton, P. Barnaghi, L. Bermudez, R. García-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson and A. Herzog, *The SSN ontology of the W3C semantic sensor network incubator group*, Web Semantics: Science, Services and Agents on the World Wide Web 17 (2012) 25-32.

[12] B. Cuenca Grau, I. Horrocks, Y. Kazakov and U. Sattler, *Modular reuse of ontologies: Theory and practice*, Journal of Artificial Intelligence Research 31 (2008) 273-318.

[13] A. Dadzie and M. Rowe, *Approaches to visualising linked data: A survey*, Semantic Web 2 (2011) 89-124.

[14] L. Daniele, F. den Hartog and J. Roes, *Created in close interaction with the industry: the smart appliances reference (SAREF) ontology*, International Workshop Formal Ontologies Meet Industries (2015) 100-112.

[15] M. d'Aquin and N. Jay, *Interpreting data mining results with linked data for learning analytics: motivation, case study and directions*, Proceedings of the Third International Conference on Learning Analytics and Knowledge (2013) 155-164.

[16] T.M. de Farias, A. Roxin and C. Nicolle, *IfcWoD, semantically adapting IFC model relations into OWL properties*, arXiv preprint arXiv:1511.03897 (2015).

[17] W. Derguech, E. Bruke and E. Curry, *An Autonomic Approach to Real-Time Predictive Analytics Using Open Data and Internet of Things*, Ubiquitous Intelligence and Computing, 2014 IEEE 11th Intl Conf on and IEEE 11th Intl Conf on and Autonomic and Trusted Computing, and IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UTC-ATC-ScalCom) (2014) 204-211.

[18] C.M. Eastman, C. Eastman, P. Teicholz, R. Sacks and K. Liston, *BIM handbook: A guide to building information modeling for owners, managers, designers, engineers and contractors*, John Wiley & Sons, 2011.

[19] R. Eastman, C. Schlenoff, S. Balakirsky and T. Hong, *A sensor ontology literature review*, 2013.

[20] V.L. Erickson and A.E. Cerpa, *Occupancy based demand response HVAC control strategy*, Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building (2010) 7-12.

[21] A. Fawzy, H.M. Mokhtar and O. Hegazy, *Outliers detection and classification in wireless sensor networks*, Egyptian Informatics Journal 14 (2013) 157-164.

[22] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, *From data mining to knowledge discovery in databases*, AI magazine 17 (1996) 37.

[23] C. Fürber, D*ata quality management with semantic technologies*, Springer, 2015.

[24] C. Fürber and M. Hepp, *Using semantic web resources for data quality management*, International Conference on Knowledge Engineering and Knowledge Management (2010) 211-225.

[25] C. Fürber and M. Hepp, *Using SPARQL and SPIN for data quality management on the semantic web*, International Conference on Business Information Systems (2010) 35-46.

[26] U. Garciarena Hualde, *An investigation of imputation methods for discrete databases and multi-variate time series*, Master's Thesis, (2016).

[27] P.A. González and J.M. Zamarreno, *Prediction of hourly energy consumption in buildings based on a feedback artificial neural network*, Energy and Buildings 37 (2005) 595-601.

[28] A. Gyrard, C. Bonnet, K. Boudaoud and M. Serrano, *LOV4IoT: A second life for ontology-based domain knowledge to build Semantic Web of Things applications*, IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud) (2016) 254-261.

[29] H. Hagras, I. Packharn, Y. Vanderstockt, N. McNulty, A. Vadher and F. Doctor, *An intelligent agent based approach for energy management in commercial buildings*, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2008) (2008) 156-162.

[30] J. Han, J. Pei and M. Kamber, *Data mining: concepts and techniques*, Elsevier, 2011.

[31] M. Hilario, A. Kalousis, P. Nguyen and A. Woznica, *A data mining ontology for algorithm selection and meta-mining*, Proceedings of the ECML/PKDD09 Workshop on 3rd generation Data Mining (SoKD-09) (2009) 76-87.

[32] V.J. Hodge and J. Austin, A survey of outlier detection methodologies, Artificial Intelligence Review 22 (2004) 85-126.

[33] M. Hofmann and R. Klinkenberg, *RapidMiner: Data mining use cases and business analytics applications*, CRC Press, 2013.

[34] R.J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*, OTexts, 2014.

[35] K. Janowicz, F. Van Harmelen, J.A. Hendler and P. Hitzler, *Why the data train needs semantic rails*, AI Magazine (2014).

[36] M. Kolchin, N. Klimov, A. Andreev, I. Shilin, D. Garayzuev, D. Mouromtsev and D. Zakoldaev, *Ontologies for Web of Things: A Pragmatic Review*, in:Anonymous , Knowledge Engineering and Semantic Web, Springer, 2015, pp. 102-116.

[37] V. Kotu and B. Deshpande, *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*, Morgan Kaufmann, 2014.

[38] F. Lécué, R. Tucker, V. Bicer, P. Tommasi, S. Tallevi-Diotallevi and M. Sbodio, *Predicting severity of road traffic congestion using semantic web technologies, in:Anonymous , The Semantic Web: Trends and Challenges*, Springer, 2014, pp. 611-627.

[39] S. Liaw, A. Rahimi, P. Ray, J. Taggart, S. Dennis, S. de Lusignan, B. Jalaludin, A. Yeo and A. Talaei-Khoei, *Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature*, International journal of medical informatics 82 (2013) 10-24.

[40] J. Lu, T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field and K. Whitehouse, *The smart thermostat: using occupancy sensors to save energy in homes*, Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems (2010) 211-224.

[41] L. Madrazo, G. Nemirovski and A. Sicilia, *Shared vocabularies to support the creation of energy urban systems models* (2013).

[42] C. Martani, D. Lee, P. Robinson, R. Britter and C. Ratti, *ENERNET: Studying the dynamic relationship between building occupancy and energy consumption*, Energy and Buildings 47 (2012) 584-591.

[43] V. Narasimha, P. Kappara, R. Ichise and O. Vyas, *LiDDM: A Data Mining System for Linked Data*, Workshop on Linked Data on the Web. CEUR Workshop Proceedings 813 (2011).

[44] A.H. Neto and F.A.S. Fiorelli, *Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption*, Energy and Buildings 40 (2008) 2169-2176.

[45] N.F. Noy, *Semantic integration: a survey of ontology-based approaches*, ACM Sigmod Record 33 (2004) 65-70.

[46] M.V. Nural, M.E. Cotterell and J.A. Miller, *Using Semantics in Predictive Big Data Analytics*, Big Data (BigData Congress), 2015 IEEE International Congress on (2015) 254-261.

[47] L. Obrst, *Ontologies for semantically interoperable systems*, Proceedings of the twelfth international conference on Information and knowledge management (2003) 366-369.

[48] F. Oldewurtel, A. Parisio, C.N. Jones, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann and M. Morari, *Use of model predictive control and weather forecasts for energy efficient building climate control*, Energy and Buildings 45 (2012) 15-27.

[49] H. Paulheim and J. Fümkranz, *Unsupervised generation of data mining features from linked open data*, Proceedings of the 2nd international conference on web intelligence, mining and semantics (2012) 31.

[50] P. Pauwels and W. Terkaj, *EXPRESS to OWL for construction industry: Towards a recommendable and usable ifcOWL ontology*, Automation in Construction 63 (2016) 100-133.

[51] P. Pauwels, S. Zhang and Y. Lee, *Semantic web technologies in AEC industry: A literature overview*, Automation in Construction (2016).

[52] Q.K. Quboa and M. Saraee, *A state-of-the-art survey on semantic web mining*, Intelligent Information Management 5 (2013) 10-17.

[53] P. Ristoski, *Towards Linked Open Data Enabled Data Mining*, in:Anonymous , The Semantic Web. Latest Advances and New Domains, Springer, 2015, pp. 772-782.

[54] P. Ristoski, C. Bizer and H. Paulheim, *Mining the web of linked data with rapidminer*, Web Semantics: Science, Services and Agents on the World Wide Web 35 (2015) 142-151.

[55] P. Ristoski and H. Paulheim, *Semantic Web in data mining and knowledge discovery: A comprehensive survey*, Web Semantics: Science, Services and Agents on the World Wide Web (2016).

[56] P. Ristoski and H. Paulheim, *Feature selection in hierarchical feature spaces*, International Conference on Discovery Science (2014) 288-300.

[57] J. Scott, A. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges and N. Villar, *PreHeat: controlling home heating using occupancy prediction*, Proceedings of the 13th international conference on Ubiquitous computing (2011) 281-290.

[58] T. Sekki, M. Airaksinen and A. Saari, *Impact of building usage and occupancy on energy consumption in Finnish daycare and school buildings*, Energy and Buildings 105 (2015) 247-257.

[59] F. Serban, J. Vanschoren, J. Kietz and A. Bernstein, *A survey of intelligent assistants for data analysis*, ACM Computing Surveys (CSUR) 45 (2013) 31.

[60] N. Seydoux, K. Drira, N. Hernandez and T. Monteil, *IoT-O, a Core-Domain IoT Ontology to Represent Connected Devices Networks*, Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016 (2016) 561-576.

[61] P.H. Shaikh, N.B.M. Nor, P. Nallagownden, I. Elamvazuthi and T. Ibrahim, *A review on optimized control systems for building energy and comfort management of smart sustainable buildings*, Renewable and Sustainable Energy Reviews 34 (2014) 409-429.

[62] A. Songpu, M.L. Kolhe, L. Jiao, N. Ulltveit-Moe and Q. Zhang, *Domestic demand predictions considering influence of external environmental parameters*, IEEE 13th International Conference on Industrial Informatics (INDIN) (2015) 640-644.

[63] P. Staroch, *A weather ontology for predictive control in smart homes*, Master's Thesis, 2013.

[64] I. Tiddi, *Explaining Data Patterns using Knowledge from the Web of Data*, The Open University (2016).

[65] P. Vandenbussche, G.A. Atemezing, M. Poveda-Villalón and B. Vatant, *Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web*, Semantic Web 8 (2017) 437-452.

[66] B.B. Wang, R.I. Mckay, H.A. Abbass and M. Barlow, *A comparative study for domain ontology guided feature extraction*, Proceedings of the 26th Australasian computer science conference-Volume 16 (2003) 69-78.

[67] Y. Wang and S. Yang, *Outlier detection from massive short documents using domain ontology*, IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS) (2010) 558-562.

[68] H. Zhao and F. Magoulès, *A review on the prediction of building energy consumption*, Renewable and Sustainable Energy Reviews 16 (2012) 3586-3592.

Appendices

A. Semantic Resources

```
PREFIX wgs84_pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX ssn: <http://www.w3.org/ns/ssn/>
PREFIX eepsa: <http://w3id.org/eepsa#>

SELECT DISTINCT ?affectingProperty
WHERE {
{?sensingdevice wgs84_pos:location <http://w3id.org/eepsa#openSpace>.
?sensingdevice eepsa:containsPhysicalObject ?sensor.
?sensor ssn:observes ?observedProperty.
<http://w3id.org/eepsa#openSpace> eepsa:isAffectedBy ?affectingProperty.
}
UNION
{
?sensor wgs84_pos:location <http://w3id.org/eepsa#openSpace>.
?sensor ssn:observes ?observedProperty
}
FILTER (?affectingProperty != ?observedProperty)
}
```

Listing 1: SPARQL query for retrieving properties that affect but are not observed within a space.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ssn: <http://www.w3.org/ns/ssn/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX eepsa: <http://w3id.org/eepsa#>
PREFIX time: <http://www.w3.org/2006/time#>
PREFIX dul: <http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#>

CONSTRUCT {?obs1 rdf:type <http://w3id.org/eepsa#OutlierCausedByRain>}
FROM <mySpace>
WHERE{
?sensdev1 rdf:type <http://www.w3.org/ns/ssn/SensingDevice>.
?sensdev1 eepsa:shelteredDevice ?sheltered.
?sensdev1 eepsa:containsPhysicalObject ?sensor1.
?sensor1 ssn:observes ?property1.
?obs1 ssn:observedBy ?sensor1.
?obs1 time:inXSDDateTime ?dateTime1.
?obs1 eepsa:obsDate ?date1.

?sensdev2 rdf:type <http://www.w3.org/ns/ssn/SensingDevice>.
?sensdev2 eepsa:containsPhysicalObject ?sensor2.
?sensor2 ssn:observes ?property2.
?obs2 ssn:observedBy ?sensor2.
```

```
?obs2 time:inXSDDateTime ?dateTime2.
?obs2 eepsa:obsDate ?date2.
?obs2 dul:hasDataValue ?val
FILTER(
<http://www.w3.org/2001/XMLSchema#boolean>(?sheltered) = false &&
(hours(<http://www.w3.org/2001/XMLSchema#dateTime>(?dateTime1))) =
(hours(<http://www.w3.org/2001/XMLSchema#dateTime>(?dateTime2))) &&
<http://www.w3.org/2001/XMLSchema#date>(?date1) =
<http://www.w3.org/2001/XMLSchema#date>(?date2) &&
?property1 = <http://purl.org/iot/vocab/m3-lite#Temperature> &&
?property2 = <http://purl.org/iot/vocab/m3-lite#Precipitation> &&
<http://www.w3.org/2001/XMLSchema#integer>(?val) > 0.2
)
}
```

Listing 2: SPARQL rule for detecting temperature outliers caused by rain.

```
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX ssn: <http://www.w3.org/ns/ssn/>
PREFIX eepsa: <http://w3id.org/eepsa#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>

SELECT ?stationID ?stationName ?city ?owner
(bif:st_distance((bif:st_point(<http://www.w3.org/2001/XMLSchema#float>(?lat),
<http://www.w3.org/2001/XMLSchema#float>(?lon))),
(bif:st_point(<http://www.w3.org/2001/XMLSchema#float>(43.19),
<http://www.w3.org/2001/XMLSchema#float>(-2.45))))) AS ?distanceToBuilding

FROM <http://tekniker.es/euskalmetStations>
WHERE {?weatherStation rdf:type "http://w3id.org/eepsa#WeatherStation".
?weatherStation foaf:name ?stationName.
?weatherStation geo:latitude ?lat.
?weatherStation geo:longitude ?lon.
?weatherStation eepsa:city ?city.
?weatherStation foaf:province ?prov.
?weatherStation foaf:owner ?owner.
?weatherStation dc:Identifier ?stationID.
?weatherStation eepsa:hasDeployed ?device.
?device eepsa:containsPhysicalObject ?sensor.
?sensor ssn:observes ?property
FILTER (
?property = "http://w3id.org/eepsa#OutdoorTemperature")
}
ORDER BY ?distanceToBuilding
```

Listing 3: GeoSPARQL query for retrieving IK4-TEKNIKER building nearby weather stations measuring temperature.