Automated Semantic Knowledge Acquisition from Sensor Data

Frieder Ganz, *Member, IEEE,* Payam Barnaghi, *Senior Member, IEEE,* and Francois Carrez

***Abstract*—The gathering of real world data is facilitated by many pervasive data sources such as sensor devices and smart phones. The abundance of the sensory data raise the need to make the data easily available and understandable for the potential users and applications. Using semantic enhancements is one approach to structure and organise the data and to make it processable and interoperable by machines. In particular, ontologies are used to represent information and their relations in machine interpretable forms. In this context, a significant amount of work has been done to create real world data description ontologies and data description models, however little effort has been done in creating and constructing meaningful topical ontologies from vast amount of sensory data by automated processes. Topical ontologies represent the knowledge from a certain domain providing a basic understanding of the concepts that serve as building blocks for further processing. There is a lack of solutions that construct the structure and relations of ontologies based on Real World Data. To address this challenge we introduce a knowledge acquisition method that processes Real World Data to automatically create and evolve topical ontologies based on rules that are automatically extracted from external sources. We use an extended k-means clustering method and apply a statistic model to extract and link relevant concepts from the raw sensor data and represent them in the form of a topical ontology. We use a rule-based system to label the concepts and make them understandable for the human user or semantic analysis and reasoning tools and software. The evaluation of our work shows that the construction of a topological ontology from raw sensor data is achievable with only small construction errors.**

I. Introduction

THERE is an emerging trend to use network-enabled devices that observe and measure the physical world to communicate the sensory data over the Internet. This growing trend towards integrating the real world data into the Internet, which is supported by Wireless Sensor Networks (WSN), RFID, smart phones, GPS device and many other sensory sources that capture and communicate the real world data is refered as *Internet of Things* (IoT). IoT defines a framework where billions of devices produce and exchange data related to real world objects (i.e. Things).

In IoT and other research domains such as pervasive and ubiquitous computing, sensor devices are often used to obtain new insights about our surrounding world and to facilitate the interaction with it. However, the increasing use of data-producing devices leads to a deluge of sensory data that requires new methods to structure and represent the information and to make the data accessible and processable for the application and services that use these data.

The semantic technologies have been used in the recent years as one of the key solutions to provide formalised representations of the real world data [1]. The advantage of applying semantic technologies to sensor data is conceptualisation and abstract representation of the raw data and making them machine interpretable and interlinking the data with existing resources on the Web.

For instance, instead of communicating and representing raw numerical values of a measurement of a weather condition, it is more desirable to use semantic concepts and properties such as *BlizzardCondition*, *isHighWindCondition* or *isFreezingCondition* measured and conceptualised by meteorological sensors.

Nevertheless, how the relationship between the raw data and its intended concept and/or relationship is established, still remains one of the biggest challenges in the IoT domain [2]. This issue is usually referred to as the symbol grounding problem [3] which describes the fundamental challenge of defining concepts from numerical sensor data that is not grounded in meaningful real world information.

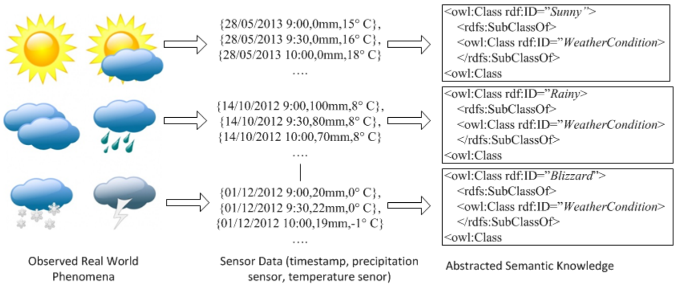
The real world data is commonly gathered as numerical values that cannot easily be related to meaningful information without knowing the context of the data such as observation time and location. The data also underlies a natural volatility and the meaning of it can change over time or it can be dependent on other external factors. For instance, 30 Celsius in summer can be a normal condition, however in winter it could be an error or an outlier condition.

In order to overcome the grounding problem, we introduce a rule-based system that designates the relationships between discretised symbolised data and semantic concepts. We provide an automated approach for a real world data driven topical ontology construction that represents the perceptual view of the collected data and relationships between different concepts. We have developed a k-means clustering algorithm to group similar discretised patterns that later represent named concepts in our ontology.

To discover relations between related concepts, we use a Markov chain model approach to find the most frequent temporal occurrences between patterns and name them after their occurrence. To label unknown concepts and properties, we use pre-defined rules. But in order to have a automated system we need an independent rule system that mines rules and applies them automatically to the concepts. To demonstrate the feasibility of our model, we introduce an approach that

Initial Manuscript submitted September 24, 2013; Major Revision submitted January 09, 2014, Minor Revision submitted February 26, 2014

Frieder Ganz, Payam Barnaghi and Francois Carrez are with the Centre for Communication Systems Research at the University of Surrey, Guildford, United Kingdom. Email: F.Ganz@Surrey.ac.uk.



```
{28/05/2013 9:00,0mm,15° C},
{28/05/2013 9:30,0mm,16° C},
{28/05/2013 10:00,0mm,18° C}
....
```

```
<owl:Class rdf:ID=”Sunny”>
   <rdfs:SubClassOf>
   <owl:Class rdf:ID=”WeatherCondition>
   </rdfs:SubClassOf>
<owl:Class
```

```
{14/10/2012 9:00,100mm,8° C},
{14/10/2012 9:30,80mm,8° C},
{14/10/2012 10:00,70mm,8° C}
....
```

```
<owl:Class rdf:ID=”Rainy>
   <rdfs:SubClassOf>
   <owl:Class rdf:ID=”WeatherCondition>
   </rdfs:SubClassOf>
<owl:Class
```

```
{01/12/2012 9:00,20mm,0° C},
{01/12/2012 9:30,22mm,0° C},
{01/12/2012 10:00,19mm,-1° C}
....
```

```
<owl:Class rdf:ID=”Blizzard”>
   <rdfs:SubClassOf>
   <owl:Class rdf:ID=”WeatherCondition>
   </rdfs:SubClassOf>
<owl:Class
```

Observed Real World Phenomena
Sensor Data (timestamp, precipitation sensor, temperature senor)
Abstracted Semantic Knowledge

Fig. 1: From Raw Data to Semantics

is able to automatically extract rules from existing resources on the web for the construction of new semantic concepts by using the raw sensory data.

We have implemented the above mentioned features and applied our solution to a real world data set with more than 250000 samples gathered from different sensors over one month period. We show that it is possible to create a topical ontology that represents basic concepts that can be used as a building block for further processing and enhancements. We evaluate the approach in terms of reconstruction error rate and show excerpts of the automatically constructed ontology.

The remainder of the paper is organised as follows: Section 2 describes the concepts of real world sensory data and semantic representation of the data. The knowledge acquisition process is presented in Section 3. We have shown the feasibility of our solution by implementing a prototype and provide the evaluation results in Section 4. Section 5 describes the related work. Section 6 concludes the paper and describes the future work.

II. Real World Data Processing Framework

The main objective of this work is to represent meaningful relations and extracted concepts from large amounts of sensory data from the real world data in a human and or machine interpretable format. As shown in Figure 1, real world phenomena are observed by collecting measurements from sensors and the raw data (mostly numerical) is sent to a user or a gateway where the data is further processed and represented in a meaningful semantic representation.

We provide a framework that infers knowledge from the data and constructs a topical ontology representation from the concepts that are extracted from the raw data. In this section we introduce some background knowledge about sensor data and discuss semantic representation frameworks.

A. Real World Data

Real World data is commonly reported through observation and measurement data obtained from sensory devices. Sensor data is often communicated as raw time-series data that can consist of a time stamp stating the time of measurement, device Id, and the values sensed by the sensor that is on board of the sensor nodes i.e. temperature, light, sound, presence and other relevant meta data.

The number of sensor nodes that are reporting data is constantly increasing. On the one hand the price for hardware is fallen and on the other hand day-to-day devices and appliances are equipped with more capable hardware. Due to the large number of sensor nodes and high sampling rates of sensor data, the amount of data is not bearable for many data processing algorithms. The deluge of data requires a variety of different efforts such as real-time reporting, spatial distribution and the variety of sensors and various qualities of the data for effective processing. Therefore dimension reduction techniques are usually used to reduce the number of features from a high-dimensional space to a low-dimensional representation [4].

Most common used techniques are: the discrete Fast Fourier Transformation (FFT), transforming the time-based data into the frequency domain to remove unwanted frequencies before transforming it back to the time-domain. The Principal Component Analysis (PCA), extracting a new orthogonal base to represent the original data by calculating the covariance or the Singular Value Decomposition (SVD), and the Piecewise Aggregate Approximation (PAA) and its symbolic representation, that uses averaged windows, utilised in this work. We evaluate and discuss some of these techniques in the evaluation section.

To abstract from numerical values and to create higher-level concepts from the large amount of data produced by sensor devices, we use the symbolic aggregate approximation (SAX) dimensionality reduction mechanism [5]. SAX discretises the data and generates symbolic words representing patterns from the sensor data.

Data discretisation serves as building block for many pattern and event detection algorithms. It enables to map reoccurring patterns to events even if there is variance, time shifting or different means in the data [6], [7], [8], [9]. In the current work, we use the symbolic approximate aggregation to transform time-series data into symbolized words as shown and explained in Figure 2.

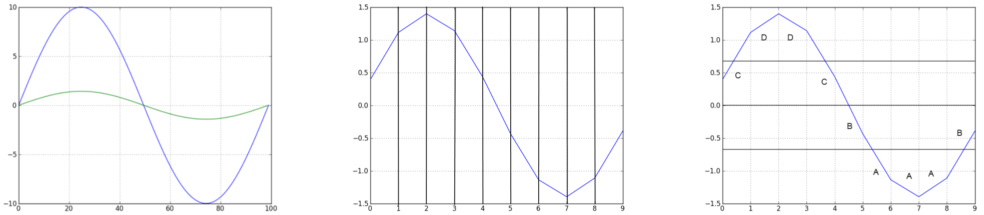
Using SAX representations, we can detect similar patterns from different sensor sources. SAX divides a window into equal segments and then creates a string representation for each segment. The SAX symbolic patterns are represented as string words, the string size and letters are adjustable in SAX (for more information refer to [8]).

For instance a time series sensor data, as shown in Figure 2, is transformed into the SAX word *“CDDCBAAAB”*; similar patterns will have resemblance to this symbolic representation. The string similarity between patterns in SAX helps to index and the compare different patterns by reducing the amount of data that has to be processed and allows to associate rules to compare and/or process the SAX words.

To illustrate the symbolic data aggregation, we use an example, the word *“CDDCBAAAB”*, is a pattern constructed from sensor data obtained from an accelerometer that has been attached to a door and measured over 5 seconds. This could lead to the semantic concept *“doorClosed”* or *“doorOpened”* that can be stored and represented in an ontology. In this work we use an extended version of the SAX algorithm, called SensorSAX.

SensorSAX is optimised for sensor data and is described in our earlier work presented in [10]. SensorSAX exploits a variable encoding rate instead of a constant rate based on the activity in the streaming data and allows higher compression

Fig. 2: Dimensionality Reduction Process of SAX



(a) The blue line represents 100 data points of a sine curve. The green (dotted) line is the curve after normalisation

(b) The normalised curve is divided into 9 windows. Each window represents the mean of 100/9 datapoints from the original data. This leads to a compression rate of n/N which is 9/100

(c) The output windows of the mean windows from (b) are divided vertical according to a Gaussian distribution. Each segment is assigned to a letter. The segment in which the curve is in per window forms the SAX word; in this case it is CDDCBAAAB

and fewer errors in reconstructing the original raw data by only transmitting SAX words in case that there is activity in the sensor data. In this work, we focus on creating a topical ontology using the patterns that are extracted from the SensorSAX patterns.

B. Semantic Representation of Real World Data

The key idea behind using semantic description for sensor data is to enable representation, formalisation and enhanced interoperability of sensor data. Ontologies can be used to store semantic concepts that represent phenomena and attributes from the real world that are understandable for the human user and also interpretable for machines due to the standardised data representation.

The concepts can be linked together through relationships that express interactions and dependencies between the concepts. The W3C Semantic Sensor Network Incubator Group has introduced the Semantic Sensor Network Ontology (SSN) [11] that provides a model to annotate sensors and their meta data, and gathered data. The SSN Ontology uses semantic concepts to model the physical attributes of sensor networks such as ”Sensor Device”, ”Temperature Sensor”, ”Radio Link”. Properties in the SSN model the relationship between concepts such as ”occuredAt”, ”observedBy” to relate sensor data annotations to domain models.

Zhao and Meersmann [12] introduce the concept of topical ontologies that represent a basic knowledge structure of a certain domain that can be used as a building block for further enhancement. Topical ontologies include the main concepts (topics) that appear in a certain domain but unlike a taxonomy also provide basic relations between the fundamental concepts. We use the SSN Ontology as a starting point for our method and extend the ontology by extracting new insights from the raw sensor data to construct a topical ontology representing a extract of the observed domain. The following describes our approach to bridge the gap between raw data and the required semantic concepts.

C. Overview of the framework

In Figure 3, an overview of the proposed framework to process the raw sensory data and construct topical ontology is shown. The framework consists of three main components: Data Pre-Processing, Ontology Construction and Rule Based Labelling. The raw sensor data serves as the input for the framework. A k-means clustering mechanism is used to group the data into clusters that form the unlabelled concepts. A Markov model is used to create temporal relations between the newly created concepts.

The unnamed concepts (i.e clustered SAX patterns) and temporal relations are used to create the initial topical ontology. After the initial ontology construction, the concepts are labelled using a rule-based reasoning mechanism. The rule-based engine processes the context of the data and tries to name the unlabelled concepts and properties. The process is shown inFigure 4 and the detailed description is found in Section 3.3.

1) **Data Pre-Processing**: In a first step, the raw data is standardised to a mean of 0 and a standard deviation of 1 to ensure an even distribution of the data over the whole processing period and allow comparison of differently distributed signals. Afterwards the data is transformed to the SAX patterns. This allows the mapping of symbolised descriptions to semantic concepts in the ontology construction and also reduces the size of data communication. The dimensionality of the data is reduced by the aggregation algorithm in SAX.
   This step can be performed on the sensing devices, in the or case that the devices are not able to perform the task due limited processing capabilities, the process can be moved to a node with higher processing capabilities (e.g a gateway).
2) **Ontology Construction**: The structure creation process defines the outline of the ontology construction. A preliminary ontology structure is created by extracting concepts and properties using a clustering algorithm and a statistical model. We follow a conceptual clustering approach [13] to create semantic concepts without labelling them.

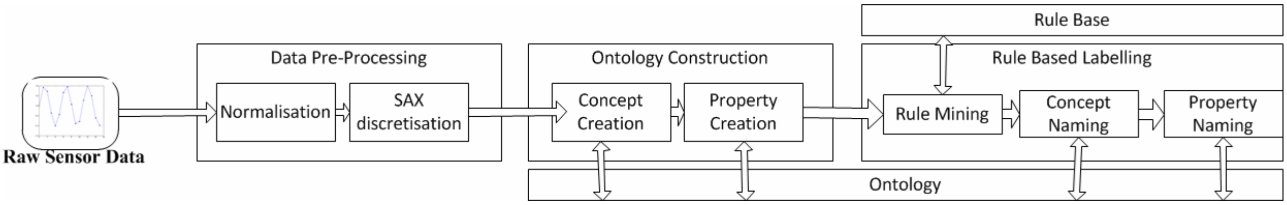


Fig. 3: Framework Overview

The clusters are formed based on the similarity of the attributes: symbolic representation and the meta-data such as sensor type and time range of the measurement. Each cluster is formalised as a unnamed concept in the ontology structure. To model the properties in our current implementation, we use a Markov model to find the temporal relations such as *"occursAfter"* between the concepts.

3) **Rule-Based Labelling**: In order to name the concepts and the properties, we utilise a rule-based mechanism. The rule system is based on the Semantic Web Rule Language. It accepts symbolised SAX patterns and adds a name tag to the unlabelled concepts.
We introduce a system that is able to extract rules based on the meta information and external data sources to automatically define the labels (this process is explained in Section 3.3).

III. Real-World Data Driven Ontology Construction

The following three methods are introduced to develop a solution that automatically constructs an ontology depicting a perceptual view of the sensed environment: clustering the symbolic patterns, creating properties via a Markov model and naming the unlabelled concepts via a rule-based method.

A. Clustering for Concept Construction

In order to reduce the amount of data that has to be processed, we use the SAX algorithm to create compressed symbolic representations of the data. SAX introduces a distance function that allows comparing generated words such as *"ABBA"* and *"ABBC"* and stating a similarity between 0 and 1. Common distance measurements and string similarity functions such as Levenshtein- or Hamming-distance cannot be used on the SAX words due to non-uniform distribution of the letters in the main SAX algorithm.

The solely comparison of the words is not sufficient, as words can be similar but measured by different type of sensors that are not related to each other. The words are also dependent on the observation time. We introduce a set of information that is needed to cluster the data into different groups based on their different attributes. We define a triple set $A = [P, t, T]$, where P is a SAX word, t is the observation type and T the observation time. In addition, we define a distance function (shown as equation 2) to compare the similarity of two triple sets;

$$saxDist(P,Q) = \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}(dist(p_i,q_i)^2} \qquad (1)$$

$$distance(A_1, A_2) = saxDist(P_1, P_2) * timeDiff(t_1, t_2) * typeDiff(T_1, T_2) \qquad (2)$$

In equation (1), $saxDist(P, Q)$ returns the distance between two words P_1 and P_1 according to the distance function in [8]. The original saxDist function is depicted in equation 1, where n is the length of the SAX word, w the alphabet size of letters used in the discretisation process and the function $dist(p_i, q_1)$ referring to a pre-calculated lookup table for the particular alphabet size w. We extend the first equation by adding a factor to compare the time difference and type difference between to triples. $timeDiff$ returning a value between]0,1] according to the temporal distance of two triples and $typeDiff$ returning either 0 or 1 matching the type of the triples. Comparing functions values from Euclidian and non-Euclidian space can lead to wrong results as the space dimensions are not equal. The alternatives are the use of non-linear dimensionality reduction techniques and the kernel trick to map them into a common space, however the complexity for sensor networks to perform this step is too high. In the evaluation section we show the feasibility of this approach.

The extracted triples from the data are fed into a k-means clustering method [14] that uses the previous defined distance function. Each cluster is transferred to an unlabelled concept that is later labelled through the rule-based mechanism. We extend the commonly used k-means algorithm for non-numerical SAX based patterns. The extended k-means algorithm is described in Algorithm 1 for two clusters.

In normal k-means, clusters are formed by calculating the distance in an Euclidean space between different sample sets. However, our sample sets consist of the non-Euclidean members: pattern, type and time. Therefore we use the distance function shown in equation 2 to measure the distances between different triples.

We start with choosing two triples randomly from the data which serve as the initial centroids for two different clusters. The distances of the other triples from the centroids are calculated according to the distance function shown in equation 2. The triples are then clustered respectively according to their distance to the centroid. The average of each cluster is then calculated with the help of the distance function and the centroids are shifted according to the average distances to the centroids. The process is repeated iteratively until both centroids converge to a certain point.

The clustering process groups similar SAX words (i.e patterns) in clusters. Each cluster is then represented as an unnamed concept. We use the clusters instead of the patterns, because small deviations in the raw data produce different patterns.

Algorithm 1 Modified K-means clustering method

```
function CLUSTERING($c1, c2$)
    for $p_i \in TriplesP$ do
        $d1 = distance(c1, p_i)$
        $d2 = distance(c2, p_i)$
        if $d1 < d2$ then
            $cluster1.add(p_i)$
        else
            $cluster2.add(p_i)$
        end if
    end for
    return $cluster1, cluster2$
end function
$c1 = random(P)$
$c2 = random(P)$
$cluster1, cluster2 = clustering(c1, c2)$
while $\epsilon \leq 0.1$ do
    $c1 = average(cluster1)$;
    $c2 = average(cluster2)$;
    $cluster1, cluster2 = clustering(c1, c2)$
    $newC1 = average(cluster1)$
    $newC2 = average(cluster1)$
    $\epsilon = (c1 - newC1) + (c2 - newC2)$
end while
```

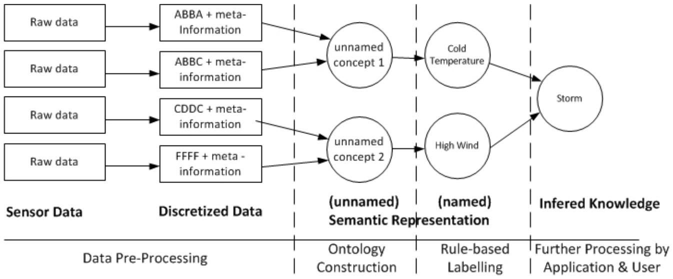


Fig. 4: Clustering the Patterns into Concepts

Using large number of patterns with small deviations to create our topical ontology will then produce a huge number of concepts and relations which provide little information. Instead we group the similar patterns and use representative concepts (i.e clusters).

B. Statistics-based property extraction

We use the frequency of unnamed concepts and their temporal occurrence to construct a Markov chain that represents the likelihood of temporal relations between unnamed concepts. The model is able to detect and represent relations between concepts through temporal properties such as *occursAfter*, *occursBefore* and *occursSame*.

We provide two options to use the extracted likelihoods. In case that the ontology supports fuzzy relations, or weighted properties as provided for instance by fuzzy OWL 2 [15] the likelihoods are noted in the semantic representation and can be used by a fuzzy reasoner [16]. If the underlying ontology or reasoner does not support fuzzy relations and only crisp and well-defined properties can be accepted, we define a threshold that has to be met to transfer the observed temporal dependency to a semantic representation. The threshold defines the level that an uncertain relation should be accepted and included in the ontology.

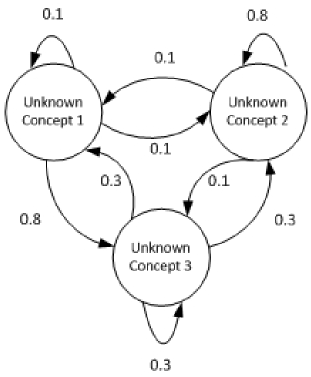


Fig. 5: Temporal Relation based on pattern frequency

It is clear that by decreasing the threshold level, more relations will be included in the ontology; however the lower threshold level will also decrease the accuracy of the property definition in the ontology. The threshold level definition is application and domain dependent and the extent of relations versus accuracy can be used to define this. The threshold can be defined using heuristics or it can be defined manually.

In this work and in the evaluation section, we show how different threshold levels affects the ontology construction. We use a heuristic defined threshold of 0.8 to filter unwanted relations. In the example shown in Figure 5, using 0.8 as the threshold level leads to defining two properties namely: ⟨ Concept 2, *"occursAfter"*, Concept 2 ⟩, ⟨ Concept 3, *"occursAfter"*, Concept 1⟩ (the direction of arrows in Figure 5 represents the temporal presents of the concepts).

C. Rule-based Concept and Property Naming

We use the Semantic Web Rule language (SWRL) [17] to label the unnamed concepts and properties that have been extracted through the clustering process. The rules defined in SWRL follow the syntax form: $antecedent \Rightarrow consequent$ where the consequent is the anticipated name to be used to label the concepts and properties and to fulfill the antecedent rules. The following shows an example of a SWRL rule:

$$isTemperatureSensor(?tmp) \& lessThanOrEqual(?tval, 8) \Rightarrow coldTemperature$$

The rules can be used to label a concept as *"coldTemperature"* in case that the antecedent conditions, "originated from a temperature sensor" and "the value of the measurement is below 8", are satisfied. In our implementation, we initially define the rules manually by considering available sensor types and possible states that can result from these types or their combination.

However, in large scale data processing scenarios manual annotation is not a feasible approach and can hinder extensibility and scalability of the solutions. There is a need to

TABLE I: Automatically extracted properties for rule creation

Sensor type	ConceptNet Attributes	Attribute	location
Temperature			
	#cold	≤5	London
	#warm	≥ 10	London
	#cool	≤10	London
	#hot	≥25	London
PIR (presence)			
	#absent	≤200	Office
	#present	≥200	Office
light			
	#light	≥100	Room
	#dark	≤100	Room
microphone			
	#soft	≤20	Room
	#loud	≥30	Room

define mechanisms that can automatically extract rules. We define a minimum set of information needed to define the rules by using sensor type, mathematical operator and location. The target is to find the labels that can be used to describe the unnamed semantic concepts. Sensor type and location information can be obtained by the sensor description that is usually available in a machine readable format.

Based on the sensor type we can infer possible attributes that can be applied as label. For this purposes we use the common sense semantic ontology "*conceptNet*" [18]. For instance we can obtain the possible labels via the sensor type *temperature*. Querying ConceptNet with *temperature* will lead to the result set *"warm"*, *"cold"*, *"hot"* and *"cool"*. By knowing the location of the sensor, we can query the ConceptNet entities and extract related attributes and causal rules relevant to a partiuclar type (e.g temperature, light). In Table I we show an excerpt of extracted properties that serve as base for the rule generation.

D. Discussion

The current work allows to create a topical ontology from raw sensor data. The created ontology can be used as a baseline for further improvements creating richer ontologies. Our approach is divided into three steps: *Data pre-processing*, *Ontology Construction* and *Rule Based labelling*.

We have attempted to propose an automated framework; however, there are certain parameters during each step that vary the outcome. In the following we describe the parameters in each step and discuss their impact. The used SAX algorithm to transform the raw numerical sensor data into string representations to reduce the dimensionality and easy comparability takes a window of samples with a specified window length and turns it into a reduced vector with a smaller lengths. The choice of the reduced vector length can have an impact on the next processing steps. In the case that a very small vector length has been chosen, important data such as outliers or certain patterns can be lost. In the case that the reduced vector length is chosen high, the effect of reducing the amount of data is decreased and either too many noise is passed onto the next algorithm or the amount of data is not suitable for the processing intensive clustering process.

Besides the parameter n, to control the reduced vector output length, a parameter a, has to be set, to control the size of the dictionary that is used when transforming to a string representation. The larger the dictionary, the more fine grain will be the resolution of the discretised representation. We have conducted some research that chooses the right parameters based on the variance of the data [19].This is useful when interesting events occur outside of the mean of a data window. Other possible techniques for the dimensionality reduction are Principal Component Analysis (PCA) and the Discrete Fourier Transformation (DFT). The different techniques are benchmarked in the evaluation section.

The Ontology Construction step uses a modified KMeans clustering algorithm that groups similar samples based on their distance. The algorithm requires two parameters, the predicted number of clusters K and a distance function. Commonly the euclidean distance is used to calculate similarity between data points, however here we use a modified distance function. In different application scenarios, variance changes in the sensor data and different scales (time function) will affect the results. To prove the feasibility a larger case-study with data-sets from more domains have to be conducted; however, this would exceed the scope of this prototypical work. There are methods to estimate the amount of clusters. In the evaluation section we introduce a method to determine the numbers of clusters based on the group variance/expected variance. There are other methods to group and classify samples that are more use-case specific such as hierarchical clustering or Mean-shift. To label the groups, we use a rule-based approach. The approach is non-parametric, and mainly relies on the knowledge base. Therefore the rule-base has to be chosen according to the application scenario. Other approaches that leverage the crowd sourced mechanisms to label the concepts are for example described in [20]

In Table II, we show the applied algorithms and summarise the used parameters that have an impact on the generated topical ontology.

IV. Implementation and Evaluation

To evaluate the proposed framework, we use sensor data obtained from the testbed in the Centre for Communication Systems Research at the University of Surrey [21]. The data is collected in one month period from sensor nodes deployed in the offices collecting information about light level, power consumption of the workstation, passive infra-red (PIR), temperature and noise levels.

We collected 274960 samples with one sample every 10 seconds and also made the data available online[1]. We first transform the data into a dimensionality reduced data set.

We have chosen to reduce the 274960 samples to a representation of 128 samples. To evaluate the reconstruction error of the different approaches, we perform the transformation over the dataset into different output vectors (with the length to the power of 2). Afterwards we reconstruct the original data from the reduced data by extrapolating[2] the data to the initial

[1] http://kat.ee.surrey.ac.uk/data.csv

[2] extrapolating by copying the values

TABLE II: Applied methods throughout the process and variable parameters

Pre Processing				
Applied Method	**Alternatives**	**Parameter**	**Description**	**Parameter Learning**
SAX	PCA, DFFT			
		n	output length	Entropy, Variance
		a	alphabet size	
Ontology Construction				
Applied Method	**Alternatives**	**Parameter**	**Description**	**Parameter Learning**
KMeans	Hiearchical Clustering, MeanShift			
		K	number of clusters	FTest, DIC
		distance	distance function	
Markov Model	Hidden Markov Model, Neuronal Network			
		t	cutoff threshold	Expectation Maximation
Labelling				
Applied Method	**Alternatives**	**Parameter**	**Description**	**Parameter Learning**
Rule-Based	Statistics, Crowdsourcing			
		Ω	Rule Base	-

c1 RC

size to be comparable. We measure the reconstruction error by taking the euclidean distance between the original data and the reconstructed data. In our approach we choose the SAX algorithm because of its simple implementation. [c1]It should be noted that the execution time of FFT is constant independently from the output size whereas SAX has a rise in execution time growing with the increasing output vector length. However, this is only valid while executed on powerful workstations where the complex computations of the FFT algorithm are supported natively SAX consists of two loops, and addition and multiplication operators in contrast to the PCA and FFT algorithms that require complex matrix computations such as eigenvalue/singular value decomposition. We argue that the complexity of the algorithm should be considered to be applicable for energy and computation constrained sensor nodes and not only measured on the common used metrics such as execution time. Especially, in heterogeneous sensor networks, where there is a large variety of different hardware, simpler algorithms ease the adaptation for different platforms. In this work the dimensionality reduction is performed on a powerful workstation, however, the dimensionality reduction process can be out-sourced to the resource constrained sensor nodes. The data is then normalised and transformed into the SAX representation to reduce its dimensionality and to make it suitable for our ontology construction algorithm. We transform each observation day into one SAX word with the length of 24 representing one letter per hour ending up with 31 words that each include 24 letters.

In a first step the data is grouped using our discrete k-means clustering method with a group size of k=2 and k=3. The k parameter is estimated by grouping the data over different k and for each k calculate the variance of each cluster group. The goal is to keep the overall variance per cluster as low as possible. In Figure 6, we show the variance over different values of k and eventually set the number of clusters to two for the following steps. The variance is used as a rule of thumb method. For future work we examine using probability density functions and data distributions to determine the number of clusters. The latter will eliminate the variance analysis and cluster number selection by using a predictive mechanism; however, the current experiment provides a better indication of the behaviour of clusters by applying different values for k but will come at a cost of accuracy. Each data point in Figure 7 represents a triple (Sax Pattern, Time stamp, Sensor type) that is grouped into one particular group. The clusters are then represented as a concept. We know from the data that two trends can be observed, the data from the power meter, noise and PIR sensors have high activity during workdays and remain steady over the weekend. The goal is to automatically label the unnamed concepts as either *workday* or *weekend* and represent this knowledge in the ontology.

We evaluate the results of the clustering method with real calendar information shown in Figure 8. The best case is to achieve an error rate of 0, thus all triples of the data set have been correctly grouped into either the *workday* or *weekend* group.

Due to the fact that we choose random triples as starting point for our clustering method the results in each experiment could be different. To show the performance of the algorithm in different experiments, we run the evaluation 100 times to get a comparable average, minimum and maximum of the error rate. The results are shown in Table III. In most cases the triples are categorised correctly, however, sometimes an odd starting triples is selected and all triples including the ones from the *weekend* are categorised as *weekdays* resulting in the highest error rate of 8. The outcome of the clustering method also highly depends on the underlying dataset. The Mic and PIR data sets lead to the smallest error rate regardless of the initial cluster set. However, the selection of the initial centroids and the cluster size has still place space for improvements and will be addressed in future work.

The groups are then included in a baseline ontology as unnamed concepts. The temporal relation between the concepts is extracted using the statistical model. In this scenario,

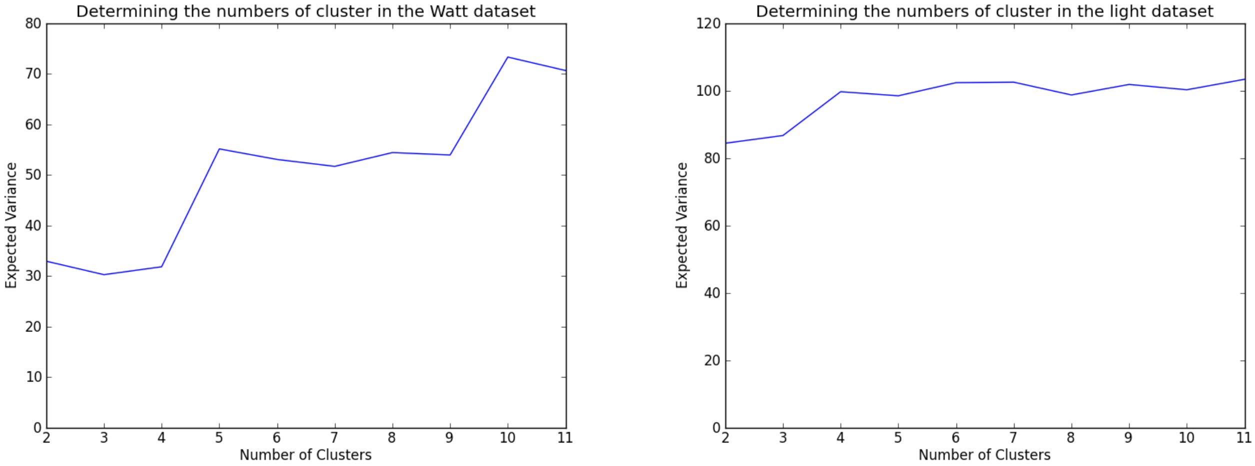


Fig. 6: Determining the numbers of clusters based on the cluster group variance

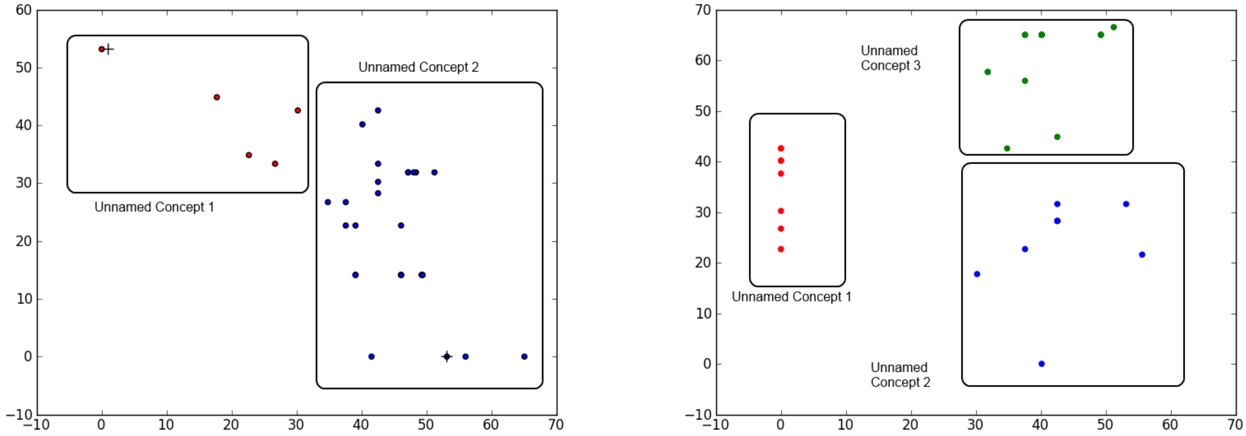


Fig. 7: Clustering the triples into groups

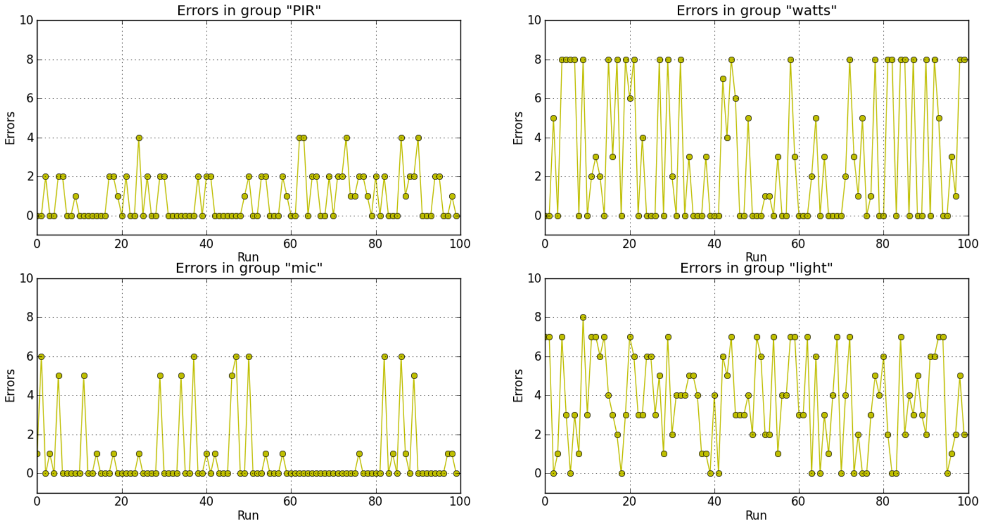


Fig. 8: Evaluating the clustered data from different sensors with real calendar information and showing the error rate over 100 random runs

TABLE III: Error Rate in detecting the correct groups from different sensor types

Error Rate	Watt	Light	PIR	Mic
Average	3.07	3.46	1.14	1.27
Minimum	0	0	0	0
Maximum	8	7	4	6

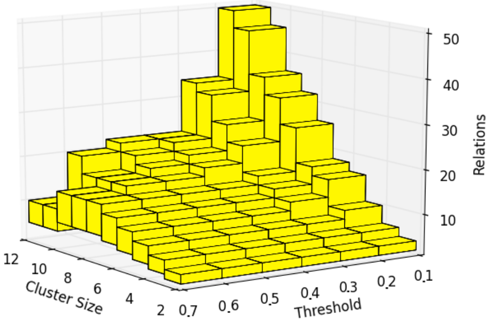


Fig. 9: Number of relations based on the factors: cut-off threshold and cluster size

it is more likely that one concept follows the same concept $p(group1 \mid group1) = 0.7$ and group 2 following after group 1 $p(group2 \mid group1) = 0.2$. This expresses that it is more likely that a *weekday* follows another *weekday*. As stated earlier the amount of (temporal) relationships is dependent on cluster size and cut-off threshold.

In Figure 9, we show the dependency between cluster size, threshold and resulting amount of relations that are eventually represented in the ontology. Currently there is no automated way to choose the right parameters and thus heuristics and domain experience have to be considered while designing the system.

Figure 10 shows an excerpt of the automatically constructed ontology. Squares represent classes that can have individuals, instances from a certain domain and represented as circles in the figure. On the left of the figure is the information that can be gathered from the sensor devices itself. Meta-information such as observation period, deployed devices and their capabilities are represented as SSN concepts. The SAX words and the inferred information that is acquired through the framework is shown on the right. The Grey highlighted concepts show the novelty of the automated process. Figure 11 shows a screenshot of the ontology visualised by an ontology visualisation tool[c0]. The framework can conclude the meaning of raw sensor data and represent it in an topological ontology. The created ontology can be downloaded[c0].

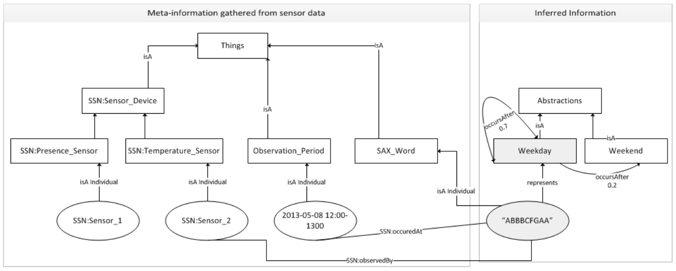


Fig. 10: A schematic view of the constructed topological ontology

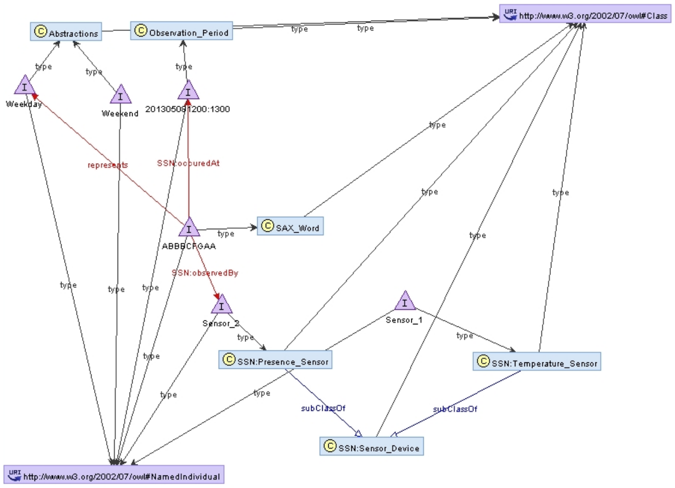


Fig. 11: An excerpt of the automatically created topological ontology

V. Related Work

In this section we provide an overview of related work on knowledge representation and creating insights from unstructured data. We also describe the methods and techniques that are used in this work.

A. Similar approaches

There are some notable work that learn and construct ontologies and semantic representations of data from text documents. Wei et al. [22] introduce an automatic learning approach to construct terminological ontologies based on different text documents. Common words that have a semantic similarity are grouped together and represented as concepts in an ontology. The authors use probabilistic models similar to our work. However, Wang et al.'s work can be applied to only text-based data.

Lin et al. [23] introduce a learning approach that constructs an ontology automatically without the requiring training data. The approach is domain-independent but also relies on text data. The keyword only data allows using existing distance metrics such as cosine distance to find similar concepts. In our approach we had to introduce a similarity function for the special domain of sensor data.

With the upcoming of the semantic sensor web, several work have been conducted in providing sensor data with semantic representations. Sheth et al. use semantics to represent and

[c0] http://semweb.salzburgresearch.at/apps/rdf-gravity/

[c0] http://http://personal.ee.surrey.ac.uk/Personal/F.Ganz/onto.owl

structure real world data in 2009 [1]. However, transforming the raw data into the semantic representation remains a challenge, especially with the deluge of data that is not not feasible to manage them using manual solutions.

Dietze et al. [24] describe the problem of symbolic grounding and the semantic sensor web and introduces an approach that uses conceptual spaces to bridge the gap between sensor measurements and symbolic ontologies in an automatic manner. The authors state that finding the right distance function remains as one of the challenges. In the current work, we define a distance function that includes the three main features of interest; pattern shape, type of measurement and observation period and provide a distance function.

Stocker et al. [25] describe a system to detect and classify different types of road vehicles passing a street with the help of vibration sensors and machine learning algorithms. The work acquires knowledge represented as an ontology by using two abstraction layers; the physical sensor layer and the sensor data layer.

However, the approach requires user interaction and/or domain-dependent training data to train the system to a certain environment. In our approach we introduce an autonomous system that does not rely on training data.

B. Data Processing mechanisms

The knowledge acquisition requires several processing steps. Due to the large volume of real world data, techniques are required to lower the amount (or dimension) of the data input to make it manageable for processing algorithms such as clustering and statistical methods. In the domain of time-series analysis there has been a large amount of dimension reduction techniques such as Fast-Fourier transformation (FFT) [26], Discrete Wavelete Transformation (DWT) [27] and Piecewise Aggregate Approximation (PAA) [28] and Symbolic Aggregate ApproXimation (SAX). The comparative study by Ding [29] reveals that SAX performs best in preserving the data features by remaining high dimension reduction (data compression).

To reduce the flood of real world sensor data we use the SAX method introduced by Lin *et al.* [8]. SAX transforms time-series data into aggregated words that can be used for pattern detection and indexing. However SAX was not developed for small constrained devices and we therefore introduce Sensor-SAX [10], a modified version that has less data transmission in times of low activity in the sensor signal that is processed.

In order to group similar types of patterns and events, clustering mechanisms are used. Cluster mechanism do not require training data and can be unsupervised. However, the clustering methods rely on distance functions that map the data samples to a comparable space. The k-means clustering method provides fast computation of the groups even in large datasets. However, the biggest drawback is that the number of clusters (i.e. k) is an input parameter, and therefore should be known beforehand.

In order to create the properties, we use a rule mining approach similar to Hu *et al.* [30]. The authors aim at creating ontologies automatically by learning the logical rules to construct the ontology. In this paper we use a rule learning approach similar to the one of Hu et al. to label the unnamed concepts in the ontology.

In overall, the concepts that are created using the clustering algorithm are named using a logical rule-mining approach. This allows the ontology construction method to insert new named concepts into the ontology and define the relations between the concepts.

VI. Conclusion

In this paper we introduce an approach that automatically generates a semantic ontology from raw sensor data that represents information and new insights gathered through a knowledge acquisition process. We introduce three mechanisms namely a discrete k-means clustering method, a statistical method and a rule-based system to provide a framework that is able to construct a knowledge representation for sensor data without requiring a preliminary training data. The proposed ontology learning solution will allow creating a machine-readable and machine-interpretable representation of the concepts (i.e. patterns) and their relations that are defined based on spatio-temporal and thematic attributes of the streams. The constructed ontology can be used in control and monitoring applications that use the sensory data to observe the status of an environment or a physical entity or it can be used to provide an overall view of the changes and related occurrences over a period.

We evaluate our approach with the help of a prototypical implementation and real world data gathered from sensors deployed in an office environment. We construct an ontology representing a work week in an office with a manual defined cluster size.

The results show an effective way in representing and naming semantic concepts from sensor data. However, the current execution time does not reflect the requirements for real-time processing, therefore more investigation has to be made to make this approach more suitable for learning and ontology construction from real-time streaming data. The approach also creates properties between semantic concepts; however, we use a pre-defined threshold to filter unwanted relationships. For the future work we will investigate how to find the best parameters for the framework in terms of window lengths for the dimensionality reduction process, the cluster size (k) and the threshold that selects the amount of relations created. Besides the technical aspects the approach can be exploited to automatically model relationships between different users and their habits (assumed the rule base contains the right information) and therefore the impact on privacy has to be considered.

Acknowledgement

This paper describes work undertaken in the context of the EU ICT Iot.est project (http://www.Ict-Iotest.eu) contract number: 288385 and EU FP7 CityPulse project (http://www.Ict-citypulse.eu) contract number: 609035.

References

[1] A. Sheth, C. Henson, and S. S. Sahoo, "Semantic sensor web," *Internet Computing, IEEE*, vol. 12, no. 4, pp. 78–83, 2008.

[2] O. Corcho and R. García-Castro, "Five challenges for the semantic sensor web," *Semantic Web*, vol. 1, no. 1, pp. 121–125, 2010.

[3] A. M. Cregan, "Symbol grounding for the semantic web," pp. 429–442, 2007.

[4] Y. Pao, "Dimension reduction, feature extraction and interpretation of data with network computing," *SERIES IN MACHINE PERCEPTION AND ARTIFICIAL INTELLIGENCE*, vol. 25, pp. 131–146, 1996.

[5] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, ser. DMKD '03. New York, NY, USA: ACM, 2003, pp. 2–11. [Online]. Available: http://doi.acm.org/10.1145/882082.882086

[6] E. Keogh, J. Lin, and A. Fu, "Hot sax: efficiently finding the most unusual time series subsequence," in *Data Mining, Fifth IEEE International Conference on*, 2005, p. 8 pp.

[7] Q. Yan, S. Xia, and Y. Shi, "An anomaly detection approach based on symbolic similarity," in *Control and Decision Conference (CCDC), 2010 Chinese*, may 2010, pp. 3003–3008.

[8] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing sax: a novel symbolic representation of time series," *Data Mining and Knowledge Discovery*, vol. 15, pp. 107–144, 2007.

[9] D. Minnen, C. Isbell, M. Essa, and T. Starner, "Detecting subdimensional motifs: An efficient algorithm for generalized multivariate pattern discovery," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, oct. 2007, pp. 601 –606.

[10] F. Ganz, P. Barnaghi, and F. Carrez, "Information abstraction for heterogeneous real world internet data," in *IEEE Sensors Journal*, vol. 13, no. 10, 2013, pp. 3793 – 3805.

[11] M. Compton, P. Barnaghi, L. Bermudez, R. García-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog *et al.*, "The ssn ontology of the w3c semantic sensor network incubator group." Elsevier, 2012.

[12] G. Zhao and R. Meersman, "Architecting ontology for scalability and versatility," in *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*. Springer, 2005, pp. 1605–1614.

[13] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Machine learning*, vol. 2, no. 2, pp. 139–172, 1987.

[14] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[15] F. Bobillo and U. Straccia, "Fuzzy ontology representation using owl 2," *International Journal of Approximate Reasoning*, vol. 52, no. 7, pp. 1073–1094, 2011.

[16] A. Emrich, F. Ganz, D. Werth, and P. Loos, "Statistics-based graphical modeling support for ontologies," in *The Seventh International Conference on Advances in Semantic Processing (SEMAPRO-13), located at NexTech 2013, September 29 - October 4, Porto, Portugal*, P. Dini, Ed. IARIA, 9 2013.

[17] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean *et al.*, "Swrl: A semantic web rule language combining owl and ruleml," *W3C Member submission*, vol. 21, p. 79, 2004.

[18] H. Liu and P. Singh, "Conceptneta practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.

[19] F. Ganz, P. Barnaghi, and F. Carrez, "Multi-resolution data communication in wireless sensor networks," in *Proceedings of the IEEE World Forum on Internet of Things*. IEEE, 2014.

[20] P. Viappiani, S. Zilles, H. J. Hamilton, and C. Boutilier, "Learning complex concepts using crowdsourcing: a bayesian approach," in *Algorithmic Decision Theory*. Springer, 2011, pp. 277–291.

[21] M. Nati, A. Gluhak, H. Abangar, and W. Headley, ". smartcampus: A user- centric testbed for internet of things experimentation," *Proc. of the 16th International Symposium on Wireless Personal Multimedia Communications*, 2013.

[22] W. Wang, P. M. Barnaghi, and A. Bargiela, "Probabilistic topic models for learning terminological ontologies," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 7, pp. 1028–1040, 2010.

[23] Z. Lin, R. Lu, Y. Xiong, and Y. Zhu, "Learning ontology automatically using topic model," in *Biomedical Engineering and Biotechnology (iCBEB), 2012 International Conference on*, 2012, pp. 360–363.

[24] S. Dietze and J. Domingue, "Bridging between sensor measurements and symbolic ontologies through conceptual spaces," 2009.

[25] M. Stocker, M. Rönkkö, and M. Kolehmainen, "Making sense of sensor data using ontology: A discussion for road vehicle classification," in *International Congress on Environmental Modelling and Software*. iEMSs, 2012.

[26] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, *Fast subsequence matching in time-series databases*. ACM, 1994, vol. 23, no. 2.

[27] K.-P. Chan and A. W.-C. Fu, "Efficient time series matching by wavelets," in *Data Engineering, 1999. Proceedings., 15th International Conference on*. IEEE, 1999, pp. 126–133.

[28] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping for datamining applications," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000, pp. 285–289.

[29] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, "Querying and mining of time series data: experimental comparison of representations and distance measures," *Proceedings of the VLDB Endowment*, vol. 1, no. 2, pp. 1542–1552, 2008.

[30] W. Hu, J. Chen, H. Zhang, and Y. Qu, "Learning complex mappings between ontologies," in *The Semantic Web*, ser. Lecture Notes in Computer Science, J. Pan, H. Chen, H.-G. Kim, J. Li, Z. Wu, I. Horrocks, R. Mizoguchi, and Z. Wu, Eds. Springer Berlin Heidelberg, 2012, vol. 7185, pp. 350–357.