Using Semantic Annotation for Knowledge Extraction from Geographically Distributed and Heterogeneous Sensor Data

Alexandra Moraru, Carolina Fortuna, Dunja Mladenić
Jožef Stefan Institute
39 Jamova, 1000 Ljubljana
Slovenia
+38614773144
firstname.lastname@ijs.si

ABSTRACT
Using semantic technologies for enriching sensor data description in scalable and heterogeneous sensor network are intended as a solution for better interoperability and easier maintainability. Through semantic annotations it is possible to provide context for sensor networks, which will improve knowledge extractions from sensor data streams and will facilitate reasoning capabilities. We propose an architecture for a system able to automatically annotate sensors descriptions, as provided by the publishers, with semantic concepts. The annotated sensor data become more meaningful and machine understandable, enabling better analysis and processing from heterogeneous streams of data. Based on the system proposed, we provide illustrative examples for demonstrating the improvements that semantic context brings and we discuss a real-world scenario of Participatory Sensing.

Categories and Subject Descriptors
I.2.4 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods – Semantic networks
H.3.4 [Information and Storage Retrieval]: Systems and Software - Question-answering (fact retrieval) systems, User profiles and alert services

General Terms
Algorithms, Experimentation, Human Factors.

Keywords

1. INTRODUCTION
The development of the Internet towards a network of interconnected objects, ranging from cars and transportation cargos to electrical appliances to any type of sensing devices, is leading to the Internet of Things. This development will provide new services and will enable new kinds of communications (i.e. “things-to-persons” or “thing-to-thing”). Furthermore, Internet of Things relies on scalable networks, mobility of wirelessly connected objects and offering interoperability for heterogeneous and complex networks [1].

Sensor Webs (or Networks) play a major role in the development of Internet of Things. In the Open Geospatial Consortium (OGC) acceptance, a Sensor Web represents a “web accessible sensor networks and archived sensor data that can be discovered and accessed using standard protocols and application program interfaces” [2], while a sensor network interconnects only sensor devices in a computer accessible network with the intention of monitoring and recording conditions at diverse location. One of the important characteristics of the Sensor Web is that its components share and use the information gathered [3]. Derived from Sensor Webs, the concept of Participatory Sensing as defined by the authors of [4], is “data collection and interpretation” and it includes mainly mobile devices that can be used to build a sensor network for capturing and sharing local data. The range of applications for this field can vary from urban planning, to public health or to natural resource management. Although the common devices that are considered to be used in Participatory Sensing are the mobile phones, the field is open to other types of sensing objects. For instance, Pachube1, one of the existing platforms for storing and sharing real-time sensor data, is enabling people to interact with sensors, from physical or virtual environments, which are connected to the internet. It registers data for over 37002 sensor nodes, with over 9400 data streams, varying from temperature, to air quality monitoring, to power consumption or to users’ Skype status. Similar projects are SensorMap3 and Sensorpedia4 also aiming at providing a “social-network” for sensors.

The principle of Participatory Sensing brings the advantage of providing access to various types of data which can be used in monitoring, studying and analyzing large scale natural and artificial systems. However, it relies on the involvement of participants or community groups for documenting the data they send. Due to the large and diverse communities that are participating at building such networks, problems may appear in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SensorKDD’10, July 25, 2010, Washington, DC, USA.
Copyright 2010 ACM 978-1-4503-0224-1...$10.00.

1 http://www.pachube.com/
2 reported on 30 April 2010
4 http://www.sensorpedia.com/
searching and finding sensors published by different participants. These problems are caused by using different vocabularies in describing the sensors and by the large and increasing number of heterogeneous sensors. Therefore, extracting knowledge from sensor description for understanding the data that it sends can be difficult, but it is also very important for maximizing the power of participatory sensing.

The documentation provided in a natural language by the participants who publish their data is not enough for a machine to understand it. A solution for improving knowledge extraction from sensor data streams is to provide semantic context. Enriching sensor description with semantic concepts leads to the development of Semantic Sensor Web (SSW), increasing interoperability and enabling complex reasoning with the contextual knowledge resulted from the semantic concepts [5].

In this paper, we propose a system architecture for semantically annotating sensor descriptions with concepts from an ontology, in order to offer a common vocabulary and a representation model which will enable better sensor discovery and will provide reasoning capabilities. Afterwards, we demonstrate with illustrative examples how the semantic context can help in complex searching of sensors and how reasoning can be applied for inferring new knowledge from the sensor descriptions. We also show the results of annotating descriptions of sensors from a real-world sensor web with semantic concepts and we discuss what improvements are required on the semantic level, without changing the design of the sensor web. To the best of our knowledge this is the first effort of annotating such a large amount of sensor data streams.

The rest of the paper is organized as follows. In Section 2 we present related work. Section 3 discusses the technologies used in semantic annotation of sensor webs and describes the system architecture that we suggest for building the SSW. Section 4 outlines the case study of participatory sensing on the Pachube platform, while conclusions and future work are included in Section 5.

2. RELATED WORK

Previous works in the Sensor Web domain have proposed and discussed different possibilities of combining Sensor Web and semantic technologies [6][7][8][9]. Through illustrative examples they explained the advantages that semantics would bring, how the resulted SSW would ease the path on using sensor data in different studies and how it will enable better communication between parties involved in building and maintaining a heterogeneous Sensor Web. In this paper we try to apply similar scenarios on a real Sensor Web, from a collaboratively environment of over 3700 sensor nodes.

In [6], the authors discuss the design of the SSW, suggesting existing data on the semantic web to be used for annotations. They present some illustrative examples for using Linked Data for annotating sensor data, exploiting the data already published. The assumption is that one is annotating sensor data directly with semantic concepts, which implies that all publishers will have to adopt a common ontology and annotate their sensor data with concepts from that ontology. An example of reasoning on semantically annotated sensor data is given, using DBpedia geographical data and then formulating a complex query in SPARQL. Our work differs in the sense that we assume that the publisher can describe the sensor data using simple tags or natural language and, afterwards, we automatically annotate those descriptions with semantic concepts.

The problem of geographical information retrieval is approached in [7]. The authors propose the use of semantic rules for adding additional processing capabilities for ontologies represented in Web Ontology Language (OWL). These rules will overtake the lack of mathematical calculus of OWL and will enable context-aware geographical information retrieval. To demonstrate the applicability of semantic rules in solving the problem, the authors developed an application ontology for the scenario of finding surf spots with respect to the users preferences. One of their achievements is that of integrating numerical data from Sensor Web with nominal data for Semantic Web, while in our work we propose a system that already incorporates this type of integration.

3. SEMANTIC ANNOTATION FOR SENSOR DATA

Extending sensor webs with semantics implies finding a suitable representation of the the afferent knowledge in such a way as to enable interoperability and reasoning mechanisms. One of the advantages that semantic technologies bring in knowledge representation are better scalability and interoperability, since adding or changing new information to a set of programs that use the same model resumes at changing the external model, while the design of those programs can remain the same, without the need of human involvement [10].

The complexity of SSW technology is derived both from the semantic and the sensor network point of view. Ontologies used in knowledge representation play a key role in usefulness of combing semantics with sensor networks. Depending on how general is the knowledge represented by an ontology they can be categorized in domain ontologies and upper ontologies. The first category represents models of specific domains (e.g. sensors) and the particular meaning of concepts related to that domain, while the second category is used to model general concepts applicable on a large set of domain ontologies. The authors of [7] are mentioning about an even more specific type of ontologies, referred to as application ontologies which “specify the conceptualization that underlie specific applications”.

Three of the existing sensor network ontologies developed are presented in [11][12][13] and have a set of common concepts related to the taxonomy of different types of sensors, physical properties of sensor devices, data acquisition and sensed domain. However, the features of the sensed domain may vary depending on the application where the sensor network is used and further development of these set of concepts is required. A detailed survey of semantic specification of sensor networks is provided in [14], where eleven sensor network ontologies are analyzed.

Such ontologies can be used for semantic annotation of sensor descriptions. Figure 1 presents the system architecture that we propose for building the SSW. We start form a sensor web composed of heterogeneous sensors which are described by their publishers. The sensor descriptions provide information about data streams, such as the type of measurements that the sensor performs (e.g. temperature, humidity, power consumption, etc.) or its physical location.

Further, there are two assumptions on which we base the rest of the system:
• the publishers provide at least a tag word for describing sensor measurements and/or location;
• the ontology concepts used for annotation are provided with a description, more exactly a string term is associated.

The next component is represented by an ontology that contains the sensor web concepts needed for annotation. These concepts are used to automatically annotate the sensor descriptions, based on the string terms associated. For demonstration purposes we utilize the Cyc [15] technology for this component and we describe it in detail in Section 3.1. Apart from the ontology, there are also the logic rules that are applied on the ontology relationships and concepts, and together with the ontology they form a knowledge base. Applying the semantic concepts on the sensor web leads to the SSW which will provide more meaningful descriptions of sensors.

The last component that we mention for our system is the inference engine, which plays an important role. The annotated sensor descriptions can be understood by such inference engines and used for solving complex queries and deriving new information. A query formulated by a user will be solved by the inference engine, which will first look into the knowledge base for the information needed for processing the query and then it will be able to search the sensors or data streams that are requested by that query, based on their annotations.

3.1 Cyc
Cyc [15] is an artificial intelligence project that aims at building a general ontology and a knowledge base for representing common sense knowledge. The Cyc technology components that present interest in this work are the knowledge base, the representation language (CycL) and the inference engine. The Cyc knowledge base is a formalized representation of fundamental human knowledge: facts, rules, and heuristics for reasoning about the objects and events of everyday life. Cyc’s knowledge is represented in CycL, while its inference engine performs general logical deduction. One of the advantages that Cyc is bringing is the very broad knowledge base covering common sense knowledge, as well as domain specific knowledge for a number of domains, which can support description of the domain of sensing for various sensor networks and also provide context for different applications. Another advantage is that of the specialized inference engine which performs modular search in the proof space enabling reasoning at large scale.

The Cyc knowledge base is organized into "microtheories", which are focused on providing context for particular domains at different level of details or different time intervals. The microtheory structure allows Cyc to independently maintain knowledge which can be contradictory for particular domains, enabling also a better performance of the system, by giving the possibility of controlling the inference process. In our work, we used the Cyc ontology without bringing any major modifications, except for introducing some simple predicates meant for illustration purposes. However, Cyc knowledge base can be modified and extended to meet the requirements of very specific domains, like the sensor web. In future work we take into consideration creating a specific microtheory that will provide the semantic context needed for building SSW.

In the rest of the paper we use CycL for formulating rules and queries, as it is an intuitive language that can be easily understood. A detailed description of CycL is beyond the scope of this paper, and it can be found in [16]. Note mentioning that Cyc concepts are represented with the #$ symbols as prefix.
3.2 Reasoning with Sensor Data

One of the most important arguments for semantically annotating sensor data is that of providing a support for performing reasoning on top of it. This will enable the possibility of applying logic rules through which new information can be inferred from the data available.

(implies
(and
(isa ?SENSOR Sensor)
(sensorMeasurementsInterval ?SENSOR ?INT)
(temporalBoundsContain ?SEASON ?INT)
(isa ?SEASON SummerSeason)
(hasRegionLocation ?SENSOR ?REGION)
(hasClimateType ?REGION MediterraneanClimateCycle)
(hasExposure ?SENSOR Outdoor)
(hasDataStream ?SENSOR ?DS)
(measures ?DS Temperature)
(valueOf ?DS (DegreeCelsius ?C))
(lessThan ?C 10))
(anomalousMeasurements ?SENSOR ?DS))

Figure 2 CycL rule for detection of anomalous measurements

An example of such new information is regarding detection of anomalous data measurements. For instance, let’s assume the following scenario:

- a large number of data stream measuring temperature in a Mediterranean region are available for summer time;
- for proper analysis of data we want to eliminate any anomalous measurements, which could have been caused by devices malfunctions or transmission errors.

Considering that there is summer season, any temperature measurements below a minimal value (e.g. 10 °C for our illustrative example) are considered anomalous for an outdoor exposure of the sensing device. The representation of such a rule in CycL can be represented as depicted in Figure 2.

A rule in CycL, begins with #Simplifies and has two parts, called its antecedent and consequent, or left-hand side and right-hand side. In our example the antecedent part is represented by all the conditions that make a data stream measurement into an anomaly. The consequent part is the assertion of a predicate which identifies the anomaly. The representation of the logic formula in natural language is:

- For every sensor S, which is performing measurements in the temporal bounds of summer season and is located in region with a Mediterranean climate, with an outdoor exposure and a data stream measuring temperature, if the measured temperature is less than 10 °C then, the sensor S is sending anomalous measurements.

When executing such a rule, inference is used also for detecting the geographical region of the sensor location based on the geographical coordinates. Similar rules can be applied for other type of measurements, such as humidity, light. Detecting an anomaly is it important not only for eliminating corrupted data before any further analysis, but also for detecting alarms, depending on the context of the problem.

4. CASE STUDY: PARTICIPATORY SENSING

An example scenario covered by the proposed system architecture is that of Participatory Sensing, which relies on publishers’ descriptions of sensors, so that the data shared can be used by others. An example of a sensor web that applies the principles of Participatory Sensing is Pachube.

<environment updated="2010-05-01T15:16:55" id="6777" >
<title> SunSPOT</title>
<feed>http://www.pachube.com/api/feeds/6777.xml</feed>
<status>live</status>
<location domain="physical" exposure="indoor">
  <name>WSN SensorLab</name>
  <lat>46.0425085163033</lat>
  <lon>14.4882792234421</lon>
</location>
<data id="0">
  <tag>Temperature</tag>
  <value minValue="18.5" maxValue="38.75" > 33.25 </value>
</data>
<data id="1">
  <tag>Light</tag>
  <value minValue="0.0" maxValue="1272.0" > 723 </value>
</data>
</environment>

Figure 3 XML description of a sensor node on Pachube

Pachube is a platform that supports storing and sharing sensor data with the aim of contributing to the building of Internet of Things. It offers support for remote environments interactions, as well as structured metadata describing the sensor data streams.

The publisher can send their data from a sensor node for storage, making it available to other users. Each sensor node has a unique id and can send more data streams from different sensor devices. When registering a sensor node on Pachube the publisher can also provide a description of the data sensed which can include information about the location of the sensor (latitude and longitude), exposure of the sensor node (indoors or outdoor), tags and units of measurements for a data stream. The descriptions are in XML format and besides the publisher data they also contain automatically generated data, such as timestamp of the last update, the current status of the sensor node (live or frozen) or minimum and maximum values for a specific stream. An example of a sensor description is presented in Figure 3.

The sensor node described in Figure 3 has the title “ SunSpot” and it has an URL address through with it can be accessed. Further, at the time of the last update (stated in the “environment” node, the “updated” attribute) the status of the sensor node was “live” meaning that it was sending data. The location node gives details about sensor location. From the description we can understand that the node is located indoor, at specified latitude and longitude coordinates. The description about the data streams are given in
the “data” node. Since a sensor node can send more data streams, each one has an id, a tag describing its measurements, last value sensed, minimum and maximum values and unit of measurements.

**Table 1. Frequent tags for data streams descriptions in Pachube**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Tag</th>
<th>Number of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature related tags</td>
<td>temperature</td>
<td>336</td>
</tr>
<tr>
<td></td>
<td>Temperature</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>temp</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>celsius</td>
<td>293</td>
</tr>
<tr>
<td>Power consumption related tags</td>
<td>electricity</td>
<td>389</td>
</tr>
<tr>
<td></td>
<td>power</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>watts</td>
<td>34</td>
</tr>
<tr>
<td>No description</td>
<td>null</td>
<td>1437</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total</th>
<th>Distinct tags</th>
<th>Data streams</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2238</td>
<td>9466</td>
</tr>
</tbody>
</table>

A particular aspect of these descriptions that captured our attention is the tags used for describing a data stream. These tags are introduced by the sensor publisher and they play a major role in sensor discovery, as one would use the tags when searching for a specific type of sensors. We have analyzed over 3700 sensor description which provide over 9400 data streams. The most frequent tags are presented in Table 1. It can be observed that for a single domain, like temperature monitoring, there can be several different tags that describe the data streams. This can bring difficulties in sensor discovery since a simple search by tag will not reveal all the sensors that one may be interested in.

**4.1 Using Semantics for Pachube Sensor Descriptions**

An important aspect of Participatory Sensing is the description that the user provides for the data sent. In general, when tagging the data sent there is no common vocabulary that the participants use. Therefore, processing these tags for extracting knowledge is required. In our approach we used Cyc ontology for finding corresponding concepts for sensor tags.

**Table 2 Cyc Concepts for Sensor tags**

<table>
<thead>
<tr>
<th>Tag</th>
<th>Cyc Concept</th>
<th>Cyc Related Strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>temperature</td>
<td>Fever</td>
<td>Temperature</td>
</tr>
<tr>
<td>temp</td>
<td>TemporaryWorker</td>
<td>Employee</td>
</tr>
<tr>
<td>celsius</td>
<td>DegreeCelsius</td>
<td>Celsius</td>
</tr>
<tr>
<td>electricity</td>
<td>Electricity</td>
<td>Electrical power</td>
</tr>
<tr>
<td>power</td>
<td>powerRating</td>
<td>Power ratings</td>
</tr>
<tr>
<td>watts</td>
<td>Watt</td>
<td>W, Watt</td>
</tr>
</tbody>
</table>

The advantage of the Cyc ontology is that it provides string terms for concepts. For instance, for the concept of #$Temperature one of the strings associated, using the termStrings predicate is “temperature” (termStrings Temperature “temperature”). This predicate can be used to find the associated concept for a string term. However, due to the very large number of concepts from very different domains that Cyc includes, problems may appear when trying to find a concept for less explicit strings.

To illustrate this we searched into the Cyc knowledge base for concepts associate to a set of sensor tags. The results are represented in Table 2, showing that a too broad ontology can introduce noise when associating concepts to strings. For instance, for the “temp” tag, which is used to represent a temperature sensor, the concept found in Cyc ( #$TemporaryWorker) is not even related to what we were looking for. Furthermore, even for an explicit tag, like “temperature” the first concept returned for associating the string with is #$Fever and the #$Temperature concept is found only in the related concepts. However, for other tags, such as “celsius” or “electricity”, it is possible to correctly annotate them with ontology concepts.

The results that we obtained for annotating sensor tags with semantic concepts require improvements for real-world scenarios. One way to achieve these improvements could be obtained by introducing context when searching the concept related to those tags. The context can be created with ontologies specialized for sensor networks or by using microtheories in Cyc. Providing a context for concept searching will reduce the number of irrelevant concepts and will provide a better categorization of sensor types of measurements.

**Individual: IJSSensor**

- isa: Sensor
- hasDataStream: IJSSensor-Data1
- hasDomain: Physical
- hasExposure: Indoor
- latitude: (Degree-UnitOfAngularMeasure 46.0425085163033)
- longitude: (Degree-UnitOfAngularMeasure 14.4882792234421)

**Individual: IJSSensor-Data1**

- isa: DataStream
- hasUnitOfMeasurement: DegreeCelsius
- measures: Temperature

**Figure 4 Representation of a Sensor Description in Cyc**

The tags used in describing data streams are important in determining what type of measurements a sensor sends (e.g. temperature, light intensity, power consumption). Nevertheless, other important features of a sensor can be found in the XML description. For instance, the geographical location of the sensor, if it is in an indoor or outdoor environment or if it sends data from a physical or virtual environment, are important features that one should take into consideration when interested in using these data streams. An illustrative example of representing a sensor node in Cyc ontology is depicted in Figure 4, which corresponds to the XML description from Figure 3. The sensor node is represented as an individual of the #$Sensor collection and it sends a data stream with temperature measurements. Further, the geographical location in terms of latitude and longitude are represented, along with few other characteristics. We would like to mention that most of the concepts already exist in the Cyc ontology (such as #$Sensor, #$DataStream, #$Latitude, #$Longitude, #$Temperature), while we also introduced some basic predicates.
for illustrating other features that are relevant (#ShasDataStream, #ShasDomain, #ShasExposure).

4.2 Searching for Sensors
Having sensor descriptions annotated with semantic concepts can improve search capabilities. Assuming that context is provided for semantic concepts search, all sensors tagged with “temperature”, “temp” or “celsius” will be annotated with the #Temperature concept. From here, retrieving all the sensors that measure temperature is a trivial task. However, more complex queries could be stated when having semantically annotated sensors.

![Query and answers in Cyc](image)

**Figure 5 Example of a query and its answers in Cyc**

An example of a complex query derives from the possibility of inferring geographical regions from sensors’ location. For instance, the distance between a sensor node and a specific location can be calculated based on geographical coordinates. This means that new type of queries can be solved, such as: “Which are the sensors that measure temperature in Ljubljana?”. Such a question can be solved in Cyc by using the #SpatialPredicate predicate, which can calculate the distance between two locations based on their latitude and longitude coordinates. The advantage of having access to a large knowledge base that incorporates common sense knowledge is highlighted in this example, where we already have represented the concepts of cities, also with details about their location. Then, we can assume that by “sensor located in Ljubljana” it is meant a fixed distance between city coordinates and sensor coordinates. For instance, we can consider that an area of 10 kilometers from the city coordinates is in the city region. An illustration of this query and the answer provided is represented in Figure 5.

After solving a query, Cyc also provides explanation of the inference performed. For the second answer of our query example, we can observe the answer bindings for each of the query variables. These bindings are resulted from simple default assertions (such as: #Sisa #SISSSensor #$Sensor) and from complex logic rules (such as the rule for computing the distance between two locations). An illustration of the summary of answer computed bindings is represented in Figure 6 or proving the validity of the inference. We show simplified explanation of the results, because the total number of steps that the Cyc inference engine performed for solving this query is 78, and a detailed explanation is out of the scope of this paper.

![Answer bindings](image)

**Figure 6 Answer bindings**

With the current^5 setting of the Pachube platform, these types of queries are impossible to solve automatically. The only search for sensors is possible by tags or title, which present the disadvantage of not returning all the answers. For location matter they provide a world-map with the location of each sensor, but the user has to manually navigate through that map.

5. CONCLUSIONS AND FUTURE WORK
A solution for enriching sensor data streams for providing machine understandable meaning is represented by the semantic technologies. Semantic annotations can provide context for the sensor measurements and observations, transforming data streams from simple binary models into meaningful data, which can be used in further analysis. In this paper we proposed and discussed a system architecture that is able to automatically annotate, with semantic concepts, sensor description provided by publishers. We demonstrated through illustrative examples the advantages of applying reasoning mechanisms on semantically enriched sensor descriptions.

We also presented the results of applying our system in a real-world scenario, that of Participatory Sensing. One of the main conclusions after analyzing these results is that a too general ontology will not be able to successfully annotate all the descriptions with relevant concepts. This problem can be solved by using a domain specific ontology or creating context in the more general ones. However, the advantages of having access to common sense knowledge have been illustrated in the examples provided in this paper, namely extended knowledge can help in inferring geographical regions or creating more complex rules.

In our future work we plan to provide more specific context for the concepts used in sensor annotation. This will enable more accurate annotation of the sensor description and will perform better in real-world scenarios. In addition, we are considering semantic solutions for sensor composition. The virtual sensor created through composition of sensors will introduce a level of abstractness that can enable better communication between people and sensor networks or between sensor networks themselves. Furthermore, the future work will be conducted using some OWL-
based ontologies in comparison with the Cyc ontology, as well as on considering different sensor description representations, including relational databases and more standardized descriptions.

### 6. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under ENVISION - ENVIRONMENTal Services Infrastructures with ONtologies (ICT-2009-249120) and PASCAL2 Network of Excellence (ICT-NoE-2008-216886).

### 7. REFERENCES


