A Framework for Semantic Enrichment of Sensor Data

Alexandra Moraru and Dunja Mladenić

Artificial Intelligence Laboratory, J. Stefan Institute, Ljubljana, Slovenia

The increased interest in sensing the environment in which we live has led to the deployment of thousands of sensors which can measure and report its status. In order to raise the impact that sensor networks can have, improving the usability and accessibility of the measurements they provide is an important step.

The problem addressed in this paper is that of enrichment of sensor descriptions and measurements in order to provide richer data, i.e., data containing more meaning. We propose a framework for automatizing the process of semantically enriching sensor descriptions and measurements with the purpose of improving the usability and accessibility of sensor data.

Keywords: semantic web, linked data, sensor web

1. Introduction

Sensors are materials or devices which change their (conductive) properties according to a physical stimulus. These sensors can be attached to more complex devices, called sensor nodes, which can have computing and communication capabilities. More and more sensor nodes are embedded into physical objects used in everyday life, ranging from pacemakers, transportation cargos to electrical appliances. Furthermore, communication links can be established between these objects, organized into wired and wireless networks called sensor networks or sensor webs in the case when web accessibility is provided [1].

Using semantic technologies for enriching sensor descriptions and measurements in scalable and heterogeneous sensor networks are intended as a solution for better interoperability and easier maintenance. Through semantic descriptions it is possible to provide context for sensor networks, which can improve knowledge extraction from sensor data streams and facilitate reasoning capabilities.

Some of the directions adopted for achieving the integration of semantic technologies and Sensor Web are related to linked data (i.e. linked sensor data [2][3][4]), or to semantic annotation and composition of web services [5]. More general directions that can be identified in building the Semantic Sensor Web are:

- Automatically annotate and enrich sensor data, by providing semantic metadata about spatiotemporal and thematic properties.

- Publish annotated sensor data using shared vocabularies and standard schemas, in order to facilitate accessibility and enable sensor discovery.

- Apply reasoning mechanisms on semantically enriched sensor data for solving problems, such as sensor composition, event detection and network management.

The enrichment of data generally refers to adding information, annotation or additional features to the data by means of computation or by pulling information from external sources (e.g., the web, databases, etc.). Semantic enrichment of sensor data denotes the process of associating semantic tags to initial sensor descriptions and measurements. These tags represent concepts, properties and relationships from an ontology and are used to describe the metadata associated to sensor data (i.e., measurement capabilities, observed phenomena, spatial properties, etc.) [6].

Making sensor data publicly available enables the development of new and useful applications.

The methods for publishing sensor data can vary from standardized web services, such as OGC's Sensor Observation Service (SOS) to application specific methods, as the ones used by web platforms, such as Pachube[1] or Sensorpedia[2]. However, such methods require prior knowledge of the infrastructures used, while publishing semantically annotated sensor data, following the linked data principles, would enable better accessibility. Moreover, when supported for integration with existing knowledge, it would increase also the usability of published data.

Reasoning, in general, is the process of producing new beliefs from a collection of believed propositions. It is strongly related to the field of logic and in the context of ontologies the logical formalisms are provided by a family of representation languages known as Description Logic (DL) [7]. Describing sensors data using ontology terms enables reasoning mechanisms that can be used to infer new knowledge for further enrichment of data or to solve complex problems.

We propose a framework for semantic enrichment of sensor descriptions and measurements, with the purposes of automatizing the process of translating existing sensor descriptions into semantic descriptions and enabling semantic querying over sensor measurements. The primary focus of this work is on the first general direction that we mentioned above, that of annotating and enriching sensor data using semantic technologies. Next, we take into consideration linked data as a method of publishing annotated sensor data, while the aspects of applying reasoning mechanism on sensor data are briefly mentioned as possible future directions.

2. Related Work

Sheth et al. [8] propose the Semantic Sensor Web (SSW) as a solution for the problem of "too much data and not enough knowledge" that appeared with the rapid development of sensor networks. In their view, the SSW represents semantically annotated sensor data with spatial, temporal and thematic metadata, facilitat-

ing thereby advanced query and reasoning. Resource Description Framework attributes (RDFa) format is adopted as an annotation language for two demonstrative applications that are using also several Sensor Web Enablement standards. Moreover, rule-based reasoning is applied for determining specific weather conditions, such as freezing or blizzard. The idea of semantic annotation is taken further by Wei and Barnaghi [9] by using Linked Open Data (LOD) resources for annotation that brings access to knowledge already represented and eliminates the risks of creating redundant data.

A recent trend for making sensor descriptions and measurements available on the Web is to publish them on LOD cloud. The advantages and challenges of Linked Sensor Data are discussed by Keler and Janowicz [4] as a solution for better sensor data accessibility without introducing very high complexity. The paper stresses out the importance of finding the appropriate links between different datasets from LOD and proposes a semiautomatic way for generating them.

Other research in the direction of publishing sensor data as linked data include [2] and [3]. Patni et al. [2] were the first to publish a large dataset of sensor descriptions and measurements, by first representing it in Observations and Measurements (O&M) standard and then converting it to Resource Description Framework (RDF) format. The linked sensor data is using a sensor ontology schema based on the concepts from O&M and the external links are made only for the location attribute, using the Geonames dataset. Barnaghi and Presser [3] propose a platform for publishing linked sensor data following the four principles proposed by Berners-Lee [10]. The platform offers an interface for publishing linked senor data without requiring from its user a related technical background. However, the user is requested to manually enter relevant keywords that describe the sensors for obtaining a list of suggested concepts from on-line repositories.

In our work we propose methods for automatizing the translation of simple sensor descriptions into semantic sensor descriptions and for processing sensor measurements for extracting

---

[1] http://www.pachube.com/
[2] http://www.sensorpedia.com/

more meaningful values for the properties observed. We adopt a strategy of analysis of sensor measurements before building the semantic representations and we use tools capable of dealing with large amounts of sensor data.

3. Semantic Enrichment of Sensor Data

The requirements for building the SSW refer to knowledge representation, description languages and semantic reasoners.

A fundamental definition of knowledge representation is given by Davis et al. [10] as "a surrogate, a substitute for the thing itself", seen as a model. One of the categories of knowledge representation appropriate for the model required is represented by ontologies.

A detailed survey of semantic specification of sensor networks is provided in [12], where eleven sensor network ontologies are analyzed. The ontologies developed for modeling sensor networks 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 this set of concepts is required. Moreover, none of the ontologies analyzed in the survey provide means to fully describe all the concepts which might be required in a real-live scenario. However, they can provide a ground to which further extension can be added to meet the specific domain requirements.

A classification in different layers of ontologies that are used to describe the sensor network domain can be found in [13]. Gray et al. are suggesting four layers of ontologies:

- upper layer, comprising upper-level ontologies used for the interoperability between other ontologies.

- infrastructure layer, describing the information required for the infrastructure (i.e., sensor network deployment, services provided by the infrastructure, metadata about sensor streams.)

- external layer, representing concepts which are not directly related to the sensor domain, such as geographical information.

- domain layer, defining the domain concepts related to a specific scenario where the sensor networks are used (e.g., floods, landslides, oil spills, etc).

The description languages are used in representing ontologies. These are named ontology languages and are included in the larger family of formal languages. Ontology languages encode the domain knowledge and the rules used for reasoning on that knowledge.

Two of the most common used languages for knowledge representation are Web Ontology Language (OWL) and RDF. Both languages are W3C standardized and different versions are defined, presenting varying levels of expressivity. RDF is a data model based on subject-predicate-object triples and uses XML for specifying syntax. RDF Schema introduces semantics to a RDF data model; it describes concepts, such as classes, properties of classes and hierarchies of these. However, RDF and RDF Schema support a limited number of semantic primitives. The advantage of OWL is better expressivity, but sometimes with higher costs regarding efficiency and reasoning capabilities.

Baader et al. [14] presented Description Logic (DL) for ontology languages, as it can provide both well-defined semantics and powerful reasoning tools. DL models concepts, roles and individuals and the relationships between them are expressed through axioms. Based on the axioms stated in DL, new relationships can be inferred using a reasoning engine that can "deduce implicit knowledge from the explicit represented knowledge" [14]. Therefore, a reasoning engine (or semantic reasoner) is a system able to draw conclusions or to infer logical consequences by applying logic rules to a set of facts or hypothesis from a knowledge base.

For the SSW domain, we classified the exiting reasoners in 3 categories: distributed or large-scale reasoners, normal scale reasoners and reasoners for constrained resource devices.

In the first category we refer to distributed platforms that are able to process large amounts of data, usually Web data. The existing or under development systems that must be mentioned are: Marvin [15] and LarKC [16]. While the first one uses a divide-conquer-swap strategy

that assures massive scalability able to eventually reach completeness, LarKC is trading computational cost to incomplete reasoning, being intended for massive heterogeneous information.

The second category covers reasoners that can normally run on a simple desktop machine and are meant for not very large ontologies, for domain specific problems where complete reasoning is required. They could be used in relatively small sensor networks. Few examples of the exiting reasoners that we considered for this category are: Pellet[3], Racer-Pro[4], FACT++[5], Cyc[6] (the reasoning component).

For the last category we consider the reasoners that can ran on resource-constrained devices, such as sensor nodes. These types of reasoners are useful in large sensor networks, where a centralized system will not perform well any more. Currently, there are some prototype implementations, one of these based on a method for automatically composing a reasoner for the needs of particular applications [17].

3.1. Linked Data

Publishing information on the Web has already changed once with the Web's evolution. If we consider Web 1.0, the data was in a static form and the interaction between the user and the data published was mostly read-only. The evolution to the Web 2.0 put the user as the central actor for generating information, through blogs or social media sites, resulting in a read-write interaction with the data. Web 3.0, also referred to as Semantic Web, uses semantic description languages such as RDF and OWL to provide a formal description of data and knowledge. Publishing information for the Semantic Web can be done via object representation described in RDF or OWL, using structured vocabularies in the form of ontologies, or as document annotated with formal metadata describing the content of the document using annotation languages (such as RDFa).

The principles of linked data have been defined by Tim Berners-Lee in [10] as follows:

1. Use URIs as names for things.

2. Use HTTP URIs so that people can look up those names.

3. When someone looks up an URI, provide useful information, using the standards (RDF, SPARQL).

4. Include links to other URIs, so that they can discover more things.

These principles have been largely adopted in the last years by the Linked Open Data (LOD) community and several methods for publishing data following these principles have been documented [18]. Therefore, one can say that publishing the related information that describes the real world object in HTML and RDF/XML representations is straightforward once it is available.

4. Conceptual Framework

The problem that the proposed framework is addressing is that of semantic enrichment of sensor descriptions and measurements with the purpose of enabling sensor discovery for better accessibility and processing of sensor data. The enrichment of data generally refers to adding information, annotation or additional features to the data by means of computation or by pulling information from external sources (i.e. the web). One example of enrichment by computation is to generate features, such as *headache likelihood* based on the barometric pressure values and their variation. An example of enrichment by pulling data from the Web is adding tweets about the weather generated around the time the values have been measured. The enriched data is then usually further processed instead of processing the original data only.

The conceptual framework illustrated in Figure 1 is defined by the following components:

- Sensor Descriptions and Measurements

- Ontology Collection

- Enrichment Components

- Semantic Repository of Sensor Data

- Data Consumers

---

[3] Pellet, http://clarkparsia.com/pellet/

[4] RacerPro, https://www.racer-systems.com/

[5] FACT++, http://owl.man.ac.uk/factplusplus/
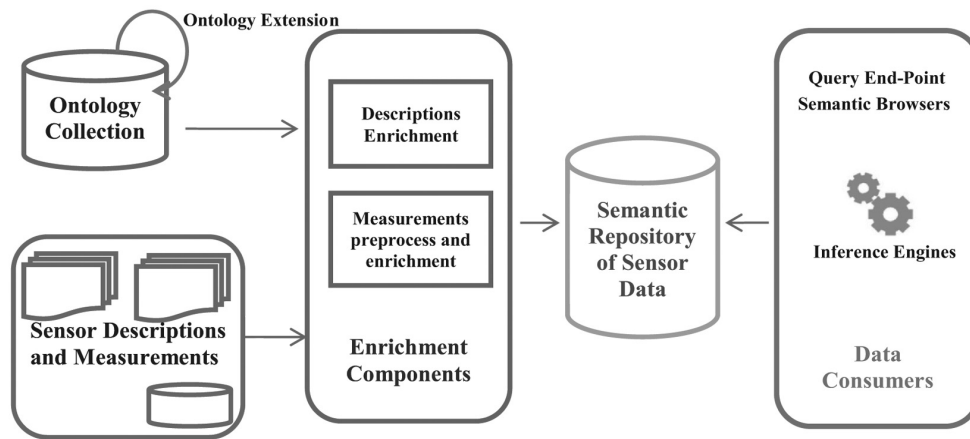
[6] Cyc, http://www.cyc.com/

Figure 1. Conceptual framework. Illustration of the main components constituting the proposed framework.

The two components to start the framework with are the Sensor Descriptions and Measurements and an Ontology Collection.

The **Sensor Descriptions** contain the metadata defining the sensor characteristics. The process of generating the metadata can follow a manual or an automatic approach. Manually generating metadata involves engineers aware of the sensor characteristics, who can describe the sensor metadata following a predefined schema (e.g. database schema, XML Schema, etc.). An automatic approach for generating sensor metadata assumes that the sensor nodes would have the capabilities of describing themselves by sending their characteristics encoded in messages to a server. In this case physical capabilities of sensor nodes are to be considered, such as memory constraints or power consumption for transmitting metadata messages. The metadata generated by either of the two methods, manually or automatically, are usually stored in databases.

The **Sensor Measurements** contain numerical values quantifying the changes of sensor properties and can be accessed using traditional database methods or streaming-based approaches.

The **Ontology Collection** consists of a set of ontologies necessary for describing sensor characteristics and providing context for sensor measurements.

The main process in the framework is run by the **Enrichment Components**, where the sensor descriptions are enriched with semantic concepts and the sensor measurements are processed to generate new features enriched with semantic concepts. The main steps of the enrichment process are:

- Analysis of the sensor descriptions and measurements for identifying the associated semantic concepts.

- Selection of the most appropriate ontologies out of the existing ontologies.

- Extension of the selected ontologies with the concepts specific to the domain of application. This mainly implies particularization of the observed properties and features of interest.

- Implementation of enrichment components, which are software programs that parse the sensor description and measurements, extracting the required metadata, and translate it with the associated semantics in a formalized language for semantic representation.

The result of the Enrichment Components involved in the framework is a **Semantic Repository of Sensor Data** (SRSD), which contains the enriched sensor descriptions and measurements.

The data from SRSD can be consumed by different **Data Consumers**, such as query end-points, semantic browsers and inference engines. The query end-point and the semantic browsers provide simple means for searching and browsing through the SRSD, supporting its representation format. The inference engines are represented by different semantic reasoners, able to infer conclusions based on the exiting facts stored in the SRSD by applying the rules defined by the ontology. A common application in which

inference engines are useful is that of virtual sensors composition, but we can also mention anomaly detection of sensor observations and sensor network management.

The conceptual components of the proposed framework can be instantiated by various implementations. As part of a master thesis [19], a deep analysis of a set of implementations has been performed together with a discussion of their advantages for different real-live scenarios.

5. Conclusions

Semantic technologies can improve the interoperability and accessibility of sensor descriptions and measurements by semantically enriching them. The context that semantic annotation provides for sensor data improves knowledge extraction and enables the development of new applications. This paper proposes a framework for enrichment of sensor descriptions and measurements that provides means for automatizing the process of semantically describing and publishing sensor data.

6. Acknowledgments

References

[1] A SEA OF SENSORS. The Economist, 2010. http://www.economist.com/node/17388356 [01/15/2012]

[2] H. PATNI, C. HENSON, A. SHETH, Linked sensor data. In *2010 International Symposium on Collaborative Technologies and Systems*, (2010 May 17-21) pp. 362–370. Chicago, USA.

[3] P. BARNAGHI, M. PRESSER, Publishing Linked Sensor Data. In *Proceedings of the 3rd International Workshop on Semantic Sensor Networks*, (2010 November).

[4] C. KELER, K. JANOWICZ, Linking Sensor data – Why, to What and How?. In *Proceedings of the 3rd International Workshop on Semantic Sensor Networks*, (2010 November).

[5] ENVISION EU PROJECT, http://www.envision-project.eu/ [01/15/2012]

[6] R. HULL, D. JENKINS, A. MCCUTCHEN, *Semantic Enrichment and Fusion of Multi-Intelligence Data.* Modus Operandi, Inc., 2009. http://www.modusoperandi.com/downloads/ Semantic_Enrichment_and_ Fusion_of_Multi_INT_Data_White_Paper.pdf [01/15/2012]

[7] R. BRACHMAN, H. J. LEVESQUE, *Knowledge Representation and Reasoning.* Elsevier, 2004.

[8] A. SHETH, C. HENSON, S. S. SAHOO, Semantic Sensor Web. *IEEE Internet Computing*, 12(4) (2008), 78–83.

[9] W. WEI, P. BARNAGHI, Semantic Annotation and Reasoning for Sensor Data. In *Proceedings of the 4th European Conference on Smart Sensing and Context*, (2009).

[10] T. BERNERS-LEE, *Design Issues – Linked Data.* http://www.w3.org/DesignIssues/ LinkedData.html [01/15/2012]

[11] R. DAVIS, H. SHROBE, P. SZOLOVITS, What is Knowledge Representation? *AI Magazine*, 14(1) (1993), 17–33.

[12] M. COMPTON ET AL., A Survey of the Semantic Specification of Sensors. In *Proceedings of the 2nd International Workshop on Semantic Sensor Networks* (2009 October).

[13] A. GRAY ET AL., A Semantically Enabled Service Architecture for Mashups over Streaming and Stored Data. In *Proceedings of Extended Semantic Web Conference 2011, Part II, LNCS 6644*, (2011) 300–314.

[14] F. BAADER, I. HORROCKS, U. SATTLER, Description Logics as Ontology Languages for the Semantic Web. *Mechanizing Mathematical Reasoning*, (2005), 228–248.

[15] E. OREN, S. KOTOULAS, G. ANADIOTIS, R. SIEBES, A. TEN TEIJE, F. MARVIN VAN HAEMELEN, A platform for large-scale analysis of Semantic Web data. In *Proceedings of International Web Science Conference*, (2009).

[16] LarKC: THE LARGE KNOWLEDGE COLLIDER. http://www.larkc.eu [01/15/2012]

[17] W. TAI, R. BRENNAN, J. KEENEY, D. O'SULLICAN, An automatically composable OWL reasoner for resource constrained devices. In *Proceedings of the 2009 IEEE International Conference on Semantic Computing*, 2009) Washington, DC, USA.

[18] T. HEATH, C. BIZER, *Linked Data: Evolving the Web into a Global Data Space.* Morgan & Claypool 2011. ISBN 978-1608454303.

[19] A. MORARU, Enrichment of Sensor Descriptions and Measurements Using Semantic Technologies. Master Thesis, Jožef Stefan International Postgraduate School, Ljubljana, 2011.

Contact addresses:
Alexandra Moraru
Artificial Intelligence Laboratory
J. Stefan Institute
Jamova 39
Ljubljana
Slovenia
e-mail: `alexandra.moraru@ijs.si`

Dunja Mladenić
Artificial Intelligence Laboratory
J. Stefan Institute
Jamova 39
Ljubljana
Slovenia
e-mail: `dunja.mladenic@ijs.si`

ALEXANDRA MORARU is a PhD student at the J. Stefan International Postgraduate School in the Information and Communication Technologies program. She got her BSc in Computer Science from the Technical University of Cluj-Napoca in 2009 and her MSc in Information and Communication Technology from the J. Stefan International Postgraduate School in 2011. She started her collaboration with the J. Stefan Institute in 2008, with a 2 months internship program, and since 2009 she is a student there. Her general research interests are in the area of semantic web and semantic technologies, more specifically, the applicability of semantics in sensor networks.

DUNJA MLADENIĆ is an expert on study and development of machine learning, data/text mining, semantic technology techniques and their application on real-world problems. She is associated with the J. Stefan Institute since 1987, first as a student and since 1992 employed as a researcher. She is leading Artificial Intelligence Laboratory of the J. Stefan Institute since 2011. She got her MSc and PhD in Computer Science from the University of Ljubljana in 1995 and 1998 respectively. She was a visiting researcher at School of Computer Science, Carnegie Mellon University, USA in 1996-1997 and in 2000-2001. She has published papers in refereed conferences and journals, served in the program committee of international conferences and organized international events in the area of text mining, link analysis and data mining. She is co-editor of several books including "Data Mining and Decision Support: Integration and Collaboration", Kluwer Academic Publishers 2003, "Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies" Springer 2008, "Web Mining: from Web to Semantic Web", Springer 2004, "Semantics, Web and Mining" Springer 2006, "From Web to Social Web: discovering and deploying user and content profiles", Springer 2007, "Knowledge Discovery Enhanced with Semantic and Social Information", Springer 2009.