

From Data to Actionable Knowledge: Big Data Challenges in the Web of Things

Payam Barnaghi, Amit Sheth, Cory Henson

Abstract— Extending the current Internet and providing connection, communication, and inter-networking between devices and physical objects, or "Things," is a growing trend that is often referred to as the Internet of Things. Integrating the real world data into the Web and providing Web-based interactions with the IoT resources is also often discussed under umbrella term of "Web of Things" (WoT). In this overview for the special issue on WoT, we describe the big data issues in the WoT, discuss the challenges of extracting actionable knowledge and insights from raw sensor data, and introduce the two papers in the special issue.

I. INTRODUCTION

The amount of data produced and communicated over the Internet and the Web is rapidly increasing. Everyday around 20 quintillion (10^{18}) bytes of data are produced (<http://www-01.ibm.com/software/data/bigdata/>). This data includes textual content (unstructured, semi-structured, structured) to multimedia content (images, video and audio), on a variety of platforms (enterprise, social media, and sensors). One of the fastest growing types of data relates to physical observations, measurements and occurrences in the real world. The growth of physical world data collection and communication is supported by low cost sensor devices such as wireless sensor nodes that can be deployed in different environments, smart phones and other network-enabled appliances. This trend will only accelerate, as it is estimated that by 2020 more than 50 billion devices will be connected to the Internet (<http://share.cisco.com/internet-of-things.html>).

Extending the current Internet and providing connections and communication between physical objects and devices, or "Things" is described under the general term of Internet of Things (IoT). Another often used term is, Internet of Everything (IOE), which recognises the key role of people or citizen sensing, such as through social media, to complement physical sensing implied by IoT. Integrating the real world data into the Web and providing Web-based interactions with the IoT resources is also often discussed under umbrella term of "Web of Things" (WoT). Data collected by different sensors and devices have various types (e.g. temperature, light, sound, and video) and are inherently diverse (the quality and validity of data can vary with different devices through time; data is also mostly location and time dependent) [10]. WoT resources can be ubiquitous and are often constrained in terms of power, memory, processing and communication capabilities. The heterogeneity, ubiquity and dynamic nature of the resources and devices, and the wide range of data, make discovering, accessing, processing, integrating and interpreting the physical world data on the Web a challenging task.

The WoT data, however, is not limited to only sensor device data. Web resident data and knowledge (e.g., Wikipedia, Linked Open Data) and information exchanged over social media and user submitted physical world observations and measurements make up a rich cyber component. Integration of physical, cyber, and social resources enables developing applications and services that can incorporate situation and context-awareness into the decision making mechanisms and can create smarter applications and enhanced services (http://wiki.knoesis.org/index.php/Smart_Data) [11], [12].

WoT data is a type of Big Data that is not only large in scale and volume, but also continuous, with rich spatiotemporal dependency. The resources that produce the data often operate in dynamic and volatile environments or can collect and communicate the data on an ad-hoc basis. The dynamicity of the environment and data providers make efficient utilisation of the WoT data on a global scale a challenging task. Figure 1 shows the access and process chain of the physical world data. The data is produced and collected using machine sensors, human sensing, smart phones and other devices. The data can be aggregated and summarised, or it can be processed and transformed to higher-level abstract descriptions of situations and events. The collected data – raw, and at times aggregated, and/or abstracted – is communicated over networks. The data can be published and stored temporarily or it can be added to repositories, and the publication interfaces can provide a metadata-enriched representation of the data. The query and discovery services can support search processes for finding the data in large-scale distributed environments. Data aggregation and summarisation can also occur at later stages by combining data from different sources and various types. The aggregation and summarisation highlight the role and importance of creating knowledge from raw data.

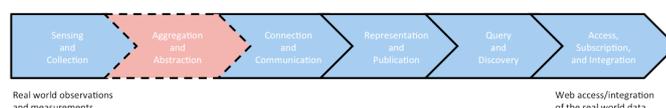


Figure 1. The data production and access chain

The most exciting outcome of addressing WoT Big Data is the new class of applications, e.g., make individualized traffic prediction and health outcomes, or more refined approaches to energy and sustainability challenges.

II. SENSING AND DATA COLLECTION

The sensor devices, smart phones, social media and citizen sensing resources are some of the key sources for producing and collecting physical world data that can be communicated,

integrated and accessed on the Web. These resources can produce large volumes of data in which the quality of the data can also vary over time. The data can be represented as numerical measurement values or as symbolic descriptions of occurrences in the physical world. Determining the quality, validity and trust of data are among the key issues in Big Data collections from the physical world, especially in use-case scenarios where the data is made available by a large number of different (and sometimes unknown) providers. As the physical world data can be related to the environment, people, and events, privacy and security are always key concerns. When the scale of the data and the number of different parties that can access and process the data increase, dealing with these issues becomes more challenging. For example, in a smart city environment where the sensory devices collect data related to citizen activities, and multiple agencies can access these data, ownership, duration of storage and types of use can raise significant privacy and security concerns.

III. CONNECTION AND COMMUNICATION

The WoT resources, especially those provided by wireless sensor devices and smart phones, have various connection interfaces and are often limited in their power and processing capabilities. It would be unrealistic to assume that all the devices will be open and available to all the requests coming from various sources on the Web. However, there is still a need for addressing and naming mechanisms to provide a means of identifying the data items and making them traceable (if required).

The large number of sensory devices, and large volumes of observation and measurement data, will require bandwidth and reliable Quality of Service solutions in order to effectively communicate these data. Different strategies, depending on the application and capabilities, are required for in-network processing, connection and middleware solutions, as well as short-term vs. long-term storage requirements and design. The pre-processing of data -- i.e., aggregation, summarisation, and/or abstraction -- can help deal with the deluge of data at the source level. Instead of surging all the raw data from sensory devices into the networks, this would allow the devices to send only higher granularity information, or digests of the data, which can be utilised in higher-level applications and services. Real-time access and mission critical applications in scenarios such as disaster monitoring and control also require efficient communication of timely data into the processing application and services from large number of distributed sources.

IV. REPRESENTATION AND PUBLICATION

Different data publishers have various ways of publishing and reporting data, and providing access to sensory data streams. Sensory data can be transient or it can be published and stored in repositories for long-term access and use. As the size and diversity of multimodal physical world data increases, publication and representation of the data in a way that makes discovery and access more flexible and scalable becomes a challenge. In recent years there have been several efforts focusing on adding enriched metadata to enhance semantic interoperability and to provide machine-readable (and potentially machine interpretable) descriptions of sensory data

– a notable example is the W3C Incubator Group on Semantic Sensor Networks [8], [4]. Several other models and semantic annotation frameworks have also been proposed for the physical world data publication and representation [2]. An important work in progress related to the representation and publication of heterogeneous physical world data includes how to automate semantic annotations, interpretation, mapping and mediation between different schema models, and efficiently balancing between express-ability and complexity of descriptions.

V. QUERY AND DISCOVERY

The query and discovery of physical world data is often based on type, time, location and the entity of interest. However, in large-scale dynamic and distributed environments, defining the region and location of requested data, indexing and querying of distributed data and/or services and data provider sources is not easy. Current solutions are very effective in processing and interpreting textual and audio-visual data. However, large scale distributed data streams that provide numerical location and time dependent data of varying quality related to physical world phenomena and discovery scenarios where the data is location and time dependent, and varies in quality, requires a different set of solutions. In principle, we still do not have fully matured search engines, similar to those on the Web, which can provide for query, indexing, discovery and resolving real-time numerical and descriptive sensory measurements and observations.

In order to better manage the tasks of publishing, sharing, analysing and understanding streaming data, researchers are adapting and extending the Semantic Web technologies. In particular, there are several efforts towards the extension of SPARQL for streaming data processing of semantically annotated data [3][1]. By extending query languages to allow continuous queries over semantically annotated data, WoT applications will more easily integrate various streams of real-world data with background domain knowledge available on the Web as Linked Open Data.

VI. ACCESS, SUBSCRIPTION, AND INTEGRATION

The physical world data is often required to be processed, analysed and interpreted in relation to other existing data on the Web. The WoT data is usually more meaningful when it is combined with metadata (e.g. what it refers to, the location and time that data is captured) and further enhanced by combining different sources or types of data to create composite and complex data types that describe a physical world phenomena or an occurrence/event related to a “Thing”. Providing automated integration and combination of data requires cooperation between various data providers that sense, measure, capture and communicate the data. It will also require flexible solutions to define composite data types, to select the resources that provide the required data and to synchronise the process. The data access scenarios can also be longer-term and continuous. In the latter case, efficient mechanisms are required to coordinate and orchestrate subscriptions to several resources for the consumers and to support mobility, access

continuity, and context-aware, energy-efficient data access and subscription to the resources.

VII. FROM DATA TO KNOWLEDGE (AGGREGATION AND ABSTRACTION)

WoT data is not only voluminous; it is also continuous, streaming, real time, dynamic and volatile. Consequently, the Big Data analytics for distributed processing of large-scale data (e.g., Hadoop) and programming models that allow automatic parallelisation of the execution of tasks (e.g., MapReduce) [7] will not be effective or adequate. In addition, creating human understandable and/or machine-readable information from raw observation and measurement data and providing real time processing and response mechanisms are also important. The distribution and efficient scalable processing of data in WoT, in addition to enhanced data publication and dissemination, will be dependent on effective mechanisms for in-network processing, aggregation and summarisation. Creating abstractions from data, or patterns of data, that can provide an aggregated view on the data will be useful. This will require using domain-specific background knowledge to extract meaningful information and actionable knowledge from the WoT data [6].

WoT resources are often dynamic; they can join a network but may later become unavailable due to network or power outage, or the source providing them can move and join a different sub-network. This will add to the challenges of discovery, integration and exploitation of data in conventional systems [9]. Indexing and discovery of resources will require a set of mechanisms that can support mobility and dynamicity in real time data and resource discovery, and can find the data by referring to their relations to objects and entities in the physical world. Unlike Internet search engines that rely on indexing existing data, the publication and integration of data in WoT cannot be separated from data discovery and search. Internet search engines discover available data, whereas in WoT data is not usually available at the time of query, and so discovery and search mechanisms would need to obtain it from suitable resources.

The quality and form of resources is another major challenge. For example, during the Fukushima disaster, when people started publishing radiation data, different users provided a wide variety of inconsistent data for similar or nearby locations [5]. Inconsistency can be due to a number of factors such as errors in reading and reporting, the use of different and un-calibrated devices, or different processes of data collection. Discovery and search methods would therefore require learning, feedback and profiling mechanisms for quality-based data queries.

In WoT, millions of devices and resources, including citizen sensors (humans reporting what they see or think using social media) participate in collecting and publishing data from the physical world. Going beyond device and resource connectivity on a large-scale, we will need data and semantic connectivity among resources and consumers for supporting the effective utilisation of the networks of the future. With the huge diversity and volumes of data expected in the near future, connectivity at the information level becomes more important

than connectivity at the network level to facilitate the effective interpretation and extraction of knowledge (i.e., abstraction) from the WoT Big Data.

Developing scalable and flexible analysis and processing models, and learning mechanisms, that can interpret large volumes of dynamic data of diverse quality requires coordination and collaboration between different methods and solutions. This includes different methods and solutions to pre-process the raw sensory data (e.g. aggregation, summarisation and filtering mechanisms), various metadata and annotation models and techniques (e.g. data representation frameworks and languages), data abstraction and pattern recognition methods, and semantic interpretation and online analytical processing methods. This will provide a value chain for the raw sensory data from various sources to be processed, integrated and interpreted, thus transformed into actionable information, insight and knowledge that lead to improved decisions and human experience. Figure 2 demonstrates different steps that can be envisaged for efficient processing and making use of the WoT data.

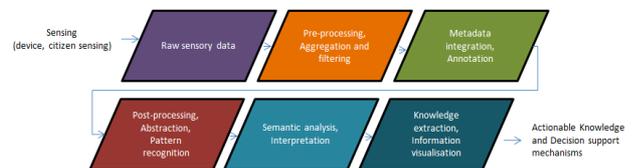


Figure 2. The process chain for the physical world data on the Web

We received 27 submissions for this special issue and identified 2 that meet high quality standards. These papers demonstrate how the WoT data can be used to monitor and interact with resources in the physical environment. In “*Farming the Web of Things*”, Kerry Taylor and her colleagues describe how sensor data can be used to monitor a Smart Farm in NSW, Australia. In the Smart Farm application, a set of environmental monitoring sensors are deployed to provide (near) real-time information related to different situations in a farm. The authors use linked-data representation of the observation data. Their proposed framework uses real-time data analysis techniques for event processing and creates semantic event descriptions that are processed to generate alerts. The paper then discusses the business challenges, barriers and drivers in using WoT technologies and sensor data in a Smart Farm environment.

In “*Human Attention Inspired Resource Allocation for Heterogeneous Sensors in the Web of Things*”, Huansheng Ning and his colleagues describe a human-attention inspired technique for resources allocation in WoT applications. The paper discusses adapting different human attention models including sustained attention, selective attention and divided attention and describes a resources allocation model that uses prior and posterior attention data to dynamically allocate resources for a WoT application.

VIII. CONCLUSIONS AND OUTLOOK

Cost efficient network-enabled devices facilitate machine-to-human and machine-to-machine communication of physical world data and its integration into the Web. Social media

platforms also facilitate publication and access to this data. The size and diversity of the generated data is growing at an extraordinary pace. The dynamicity, volatility and ad-hoc nature of most of the underlying networks and resources that produce physical world observation and measurement data introduce additional challenges for processing and utilisation of Big Data. The physical world is also often related to people and their surroundings, so ethical issues and privacy and security concerns always remain at the heart of WoT systems and applications. The limitations and status of energy and resource constrained devices and networks, and the ability to effectively publish, discover and access the data in large-scale distributed environments, have a direct impact on the performance of the systems that use this data. Semantic enhancements and metadata are also important to make the physical world data interoperable and interpretable by the automated software tools and services. Providing efficient and scalable solutions for annotation of the physical world data and offering solutions that can provide high performance and sometimes (near) real-time analytics are also crucial in the WoT systems. In conclusion, WoT is a driving force to generate, access and integrate data from physical, cyber and social sources. The ability to develop solutions that can effectively analyse and interpret physical world data requires collection and integration of data from various sources. The ability to analyse and interpret this data to create meaningful insights, extract knowledge and create situational awareness is crucial to fulfilling the potential of big Web of Things data.

REFERENCES

- [1] D. Anicic, P. Fodor, S. Rudolph, and N. Stojanovic, "EP-SPARQL: a unified language for event processing and stream reasoning", World Wide Web Conference (WWW), pp. 635–644, 2011.
- [2] P. Barnaghi, W. Wang, C. Henson and K. Taylor, "Semantics for the Internet of Things: Early Progress and Back to the Future", International Journal on Semantic Web and Information Systems (IJSWIS), 8(1), 1-21, 2012. doi:10.4018/jswis.2012010101
- [3] A. Bolles, M. Grawunder, and J. Jacobi, "Streaming SPARQL extending SPARQL to process data streams", Extended Semantic Web Conference (ESWC), pp. 448–462, 2008.
- [4] M. Compton, P. Barnaghi, L. Bermudez, R. Garcia-Castro, et al, "The SSN ontology of the W3C semantic sensor network incubator group", J. of Web Semantics, vol 17, pp. 25-32, 2012.
- [5] S. Haller, "Linked Data Use and the Internet of Things", SAP Research, Future Interent Assembly, May 2011, available at: <http://goo.gl/xI4mD>
- [6] C. Henson, K. Thirunarayan, and A. Sheth, "An Efficient Bit Vector Approach to Semantics-based Machine Perception in Resource-Constrained Devices", In Proceedings of 11th International Semantic Web Conference (ISWC 2012), Boston, Massachusetts, USA, Nov. 11-15, 2012.
- [7] T. Kraska, "Finding the Needle in the Big Data Systems Haystack", Internet Computing, IEEE , vol.17, no.1, pp.84,86, Jan.-Feb. 2013.
- [8] L. Lefort, C. Henson, K. Taylor et al, "Semantic Sensor Network XG Final Report", W3C Incubator Group Report, (2011).
- [9] H. G. Miller, P. Mork, "From Data to Decisions: A Value Chain for Big Data", IT Professional , vol.15, no.1, pp.57,59, Jan.-Feb. 2013.
- [10] A. Sheth, C. Henson, S. Sahoo, "Semantic Sensor Web", IEEE Internet Computing, vol. 12, no. 4, pp. 78-83, July-Aug. 2008.
- [11] A. Sheth, P. Anantharam, C. Henson, "Physical-Cyber-Social Computing: An Early 21st Century Approach", IEEE Intelligent Systems, pp. 79-82, Jan./Feb. 2013.
- [12] K. Thirunarayan and A. Sheth, "Semantics-empowered Approaches to Big Data Processing for Physical-Cyber-Social Applications", AAAI 2013 Fall Symposium on Semantics for Big Data, Arlington, Virginia, November 15-17, 2013.

Payam Barnaghi is with the University of Surrey. Contact him at: p.barnaghi@surrey.ac.uk; <http://personal.ee.surrey.ac.uk/Personal/P.Barnaghi>.

Amit Sheth is with Kno.e.sis at Wright State University. Contact him at amit@knoesis.org; <http://knoesis.org/amit>.

Cory Henson with Kno.e.sis at Wright State University. Contact him at cory@knoesis.org; <http://knoesis.org/researchers/cory>.