International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 3.45 (SJIF-2015), e-ISSN: 2455-2585
Volume 4, Issue 5, May-2018

Data Stream Learning in IoT Environment: A Survey

Viren Patel[1], Viral Borisagar[2]

[1]Assistant Professor, Computer Engineering Department, Government Engineering College, Dahod
virenjpatel@gmail.com
[2]Assistant Professor, Computer Engineering Department, Government Engineering College, Patan
viralborisagar@gmail.com

ABSTRACT

Smart systems need to take faster decisions. Computation of stored data induces delays and renders a real time system useless. Sensors are not smart devices, they collect data at edge level of network. Uploading the data to cloud for processing is slow, a faster way to predict from the data collected is by processing it at the lower levels of network. Machine Learning concepts can be exploited to train the collaborating devices to represent such data for information exchange. This paper discusses background of possible research areas open for proposing decision making systems at lower layers of network hierarchy.

Keywords: Stream Data Processing, Stream Learning, IoT, Fog Computing, Edge Computing, Data Abstraction, Machine Learning

1. INTRODUCTION

THINGS that collect data, act on data, can connect to other things and exchange data are growing in abundance. These devices sum up to make lowest level of large networks of intelligent things. The IoT like any new technological paradigm has brought up challenges in designing it, utilizing it, making it secure and exploiting for the best of our needs in day to day life. IoT will change the way we spend our time, make decisions and get help. Soon by 2020 we will cross the mark of generating 40 Yotta bytes of data per day [1]. The earlier IoT devices were much simpler like RFID. Today they are any small battery powered blue tooth enabled device like a wearable pedometer. We have slowly accustomed ourselves to the fact that our device can give us suggestions which are both timely and accurate most of the times. We have been using simple machine like alarm clocks to wake us up in the morning, or remind us of some important appointment, but today we have come far from making just one machine intelligent. We have lots of intelligent, communicating machines around us which are built and programmed to help us with our routines and sometimes with critical circumstances. Temporal data have been collected in various fields, such as brain science, ecology, geophysics, social sciences. Temporal data may contain complex temporal patterns that would need to be learned and extracted in a computational model.

The term Web of Things or Internet of Things (IoT) was coined much earlier before technology could use the data collected for meaningful purpose. IoT deals with data from numerous amounts of sources in different formats. We have been capable of collecting that data and run data mining algorithms like classification, clustering and association rule mining. Sensor nodes are usually scattered in a sensor field; each of these scattered sensor nodes has the capabilities to collect data and route data back to a special node called sink by a multi-hop infrastructure less architecture. Large scale initiatives are underway in Japan, Korea, the USA and Australia, where industry, associated organizations and government departments are collaborating on various programs, advancing related capabilities towards IoT. This includes smart city initiatives, smart grid programs incorporating smart metering technologies and roll-out of high speed broadband infrastructure [2]. The amount to data processed increases from layers of data collection until it reaches to cloud. [3]. The future trends point to a world wherein we will require our devices to make decisions on the fly based on various environmental parameters collected by sensors (things) and calculated by devices (things) in collaboration with or without other data from external sources with similarly deployed devices. For example, 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 isHighWindCondition or isFreezingCondition measured and conceptualized by meteorological sensors [4]. The devices at the edge of network have become data producers unlike to prior scenario when they were only data consumers [5].

2. RELATED CONCEPTS

2.1 Data Streams

Streaming Data is data that is continuously generated by different sources. Such data should be processed incrementally using Stream Processing techniques without having access to all of the data. Data Streams can be defined differently based on perspectives. The common property is its flow that is fast and unpredictable in frequency. A data stream is a potentially unbounded, ordered sequence of data items which arrive over time. The time intervals between the arrivals of each data item may vary. These data items can be simple attribute-value pairs like relational database tuples, or more complex structures such as graphs [6].

2.2 Data Ingestion

The amount of data a system can process while the data is in motion. The challenge lies in storing only the inferred information from flow of continuous stream of data.

2.3 Data Stream Learning

Applying machine learning algorithms on streams of data will give results that can predict future trends of streams. Techniques to learn from Streams of continuously flowing data have been research interest from many years now. Classification of such streams and detecting new class or deciding splitting criteria within a class is areas of research concerned with Stream Learning.

2.4 Concept Drift

Concept Drift occurs in a stream when there is a change in distribution of data. As the environment where the data are collected may change dynamically, the data distribution may also change accordingly. This phenomenon, referred to as concept drift, is one of the most important challenges in data stream mining. For example, in Indian sub-continent 36 degree Celsius is low temperature in Summer but in Winter 16 degree Celsius is low. Around the year, the concept of ’pleasant’ temperature of season keeps changing. In practice, the concept drift phenomenon is most commonly reflected in its consequences - a deteriorating prediction performance in incremental learning. The learned knowledge of the model becomes obsolete in the face of the examples with a new background concept. Finding the exact points of change can be very challenging, as the change between two distributions can be gradual and has to be differentiated from transient noise that can affect the stream [7].

2.5 Class Evolution and Ensemble Learning

Class evolution is different from concept drift. Rather than change in data distribution, it tracks the prior probability distribution of classes. [8] The posts on social media keep bringing up new topics. Introduction of each new topic leads to emergence of class. Same way the class emerged can also disappear and reappear while posts are made from time to time. Class evolution models must be explored exhaustively to exploit Data Stream learning. Extensive survey on Ensemble Learning techniques is also made [6].

2.6 Complex Event Processing

Complex Event Processing is Event Processing that combines input from multiple sources to infer events or patterns that suggest more complicated circumstances. The CEP must identify meaningful events, such as threat or opportunity and respond to them as quickly as possible.

2.7 Dimensionality Reduction

In statistics, machine learning, and information theory, dimensionality reduction or dimension reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. It can be divided into feature selection and feature extraction. In order to avoid the effects of the curse of dimensionality, dimension reduction is usually performed prior to applying a Knearest neighbors algorithm (k-NN) to high dimensional datasets. Feature extraction and dimension reduction can be combined in one step using principal component analysis (PCA), linear discriminant analysis (LDA), canonical correlation analysis (CCA), or non-negative matrix factorization (NMF) techniques as a preprocessing step followed by clustering by KNN on feature vectors in reduced-dimension space. In machine learning this process is also called low-dimensional embedding.

2.8 Ontology

Ontology is the representation of entities, ideas and events, along with their properties and relations, according to a system of categories. Ontology learning is the automatic or semi-automatic creation of ontologies, including extracting a domain’s terms from natural language text. As building ontologies manually is extremely labor-intensive and time consuming, there is great motivation to automate the process. Information extraction and text mining have been explored to automatically link ontologies to documents, for example in the context of the BioCreative challenges.

2.9 Semantic Web

The Semantic Web is an extension of the World Wide Web through standards by the World Wide Web Consortium (W3C).The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF). According to the W3C, The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries [2]. The Semantic Web is therefore regarded as an integrator across different content, information applications and systems.

2.10 Semantic Sensor Web

Sensor networks provide the challenge of too much data, and too little inter-operability and also too little knowledge about the ability to use the different resources which are available in real time. The Sensor Web Enablement initiative of the Open Geospatial Consortium defines service interfaces which enable an interoperable usage of sensor resources by enabling their discovery, access, tasking, eventing and alerting [9]. Advanced machine learning (ML) techniques show the potential to automate a number of relevant tasks for the Semantic Web (SW) by complementing and integrating logical inference with inductive procedures that exploit regularities in the data. An obvious benefit of inductive methods is that they are more robust against some of the inherent problems of the SW such as contradicting information, incomplete information and non-stationary [10].

2.11Edge Computing

Edge computing is enabling technologies allowing computation at the edge of network [5]. By edge we mean the path for upstream data from data source to cloud and for downstream data from cloud to data source. This makes devices like mobile phone an edge level device. Same analogy applied to large system implementing sensors for collecting any geographical data can have edge devices that collect and pre-process data before uploading them to upper layers. Making these devices do the computation is the area of research. Edge computing is the new platform for computing in large-scale geospatially distributed and latency sensitive networks [3].

2.12Fog Layer

A computing layer that sits between the edge devices and the cloud in the network topology is conceptualized as Fog layer, the processing to be designed at this level can be called fog computing. They have more compute capacity than the edge but much less so than cloud data centers. They typically have high uptime and always-on Internet connectivity. Applications that make use of the fog can avoid the network performance limitation of cloud computing while being less resource constrained than edge computing. As a result, they offer a useful balance of the current paradigms [11].

3. Literature Survey

It is worth noting that considerable research is made in developing systems that are light weight and fast in decision making. Since the introduction of term IoT the advancement in technology has increased use of sensor based devices. This in turn will increase the amount of data collected at aggregation layer. All this data being unstructured and varied in format and specification; we can call it big data. Survey on impact of IoT and its possible applications have found following areas to be most affected in near future: Transportation and Logistics, Healthcare, Smart environments, Smart cities and Smart locomotives [12]. In one such survey [13] the authors propose a data management framework for huge amount of data generated by IoT.

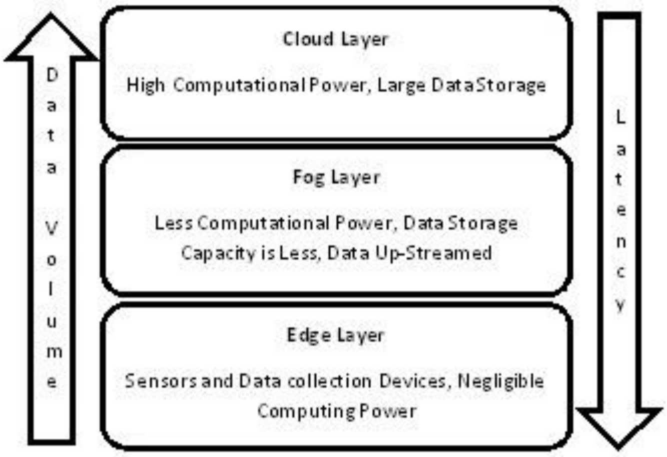
Data Volume
Cloud Layer
High Computational Power, Large DataStorage
Fog Layer
Less Computational Power, Data Storage Capacity is Less, Data Up-Streamed
Edge Layer
Sensors and Data collection Devices, Negligible Computing Power
Latency

Figure 1 Layer Concept

Data Stream learning involves major contribution of classification techniques to be implemented on flowing data. A new splitting criteria for such Stream Learning scenarios is presented in [14]. Extensive research has been made in recent years in the area of ensemble learning from data streams [6]. Data Stream Learning based on concept drift has been discussed in [7]. Drifts can be made adaptive by adjusting cluster mean as data changes. Adaptive clustering technique is proposed in [15]. In [16] a semi-supervised framework for classifying evolving data streams is proposed. It detects concept drift and determines chunk boundary dynamically by finding any significant change in classifier confidence. Moreover, it also uses confidence scores to intelligently select limited amount of data instances for labeling from the latest chunk, which is then used to update the classifier.

Edge computing will have impact on implementations of Cloud Offloading, Video Analytics, Smart Home and Smart Cities [5]. The data accumulated at edge of IoT Environment will be unformatted, unstructured and huge in amount, making it big data. To address such challenges the Semantic Web and its derivatives in the form of Linked data and Web of data can play a crucial role [1].

We found many solutions proposed for targeted environments where streams were captured to learn and predict. A predictive maintenance application presented in [17] captures sound data streams to predict wear and tear in machines. The edge of IoT environment will have data producing devices. Only the layer above this edge can be used for smart applications. In [12] the middle layer of IoT environment is shown to imbibe object abstraction, service composition, trust and privacy management and IoT applications.

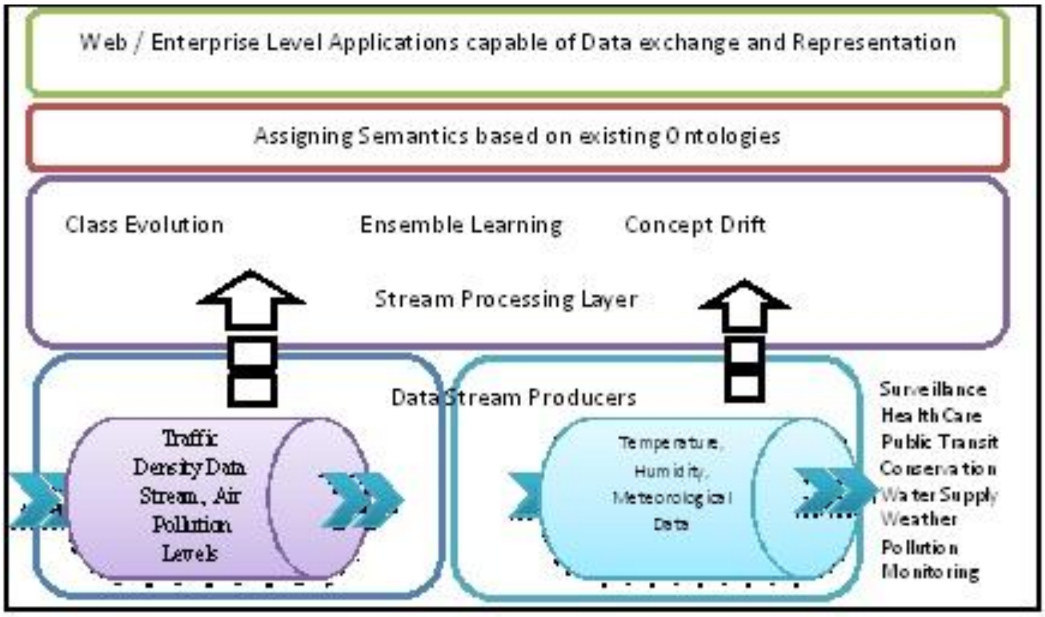
Web / Enterprise Level Applications capable of Data exchange and Representation
Assigning Semantics based on existing Ontologies
Class Evolution
Ensemble Learning
Concept Drift
Stream Processing Layer
Data Stream Producers
Traffic Density Data Stream, Air Pollution Levels
Temperature, Humidity, Meteorological Data
Surveillance
Health Care
Public Transit
Conservation
Water Supply
Weather
Pollution Monitoring

Figure 2 Semantics for Moving Data

Recent research consider constraints of edge layer devices in terms of computing power and thus direct research on aggregating data collected to semantics. These semantics either new or extension of already developed libraries for SSN(Semantic Sensor Networks) . Big data is more valuable when we can make quick analysis [18]. A network of networks of sensory devices is basic assumption for any smart decision making system. Discussion on smart cities elicits different architectural requirements for multi-layered network [19]. Decision making at cloud level is discouraged as its latency will not permit timely decision making. Pushing the goal to lower levels, a new decision making layer is proposed. This fog layer is above the edge of network that contains sensors [3]. An anticipatory learning model has been presented in [3] that suggest utilizing sensor data at edge level and up streaming it to fog layer for contextualization and further up-streaming to cloud for predicting next data values of sensors at edge layer. They have presented an anticipatory behavioral pattern on a transit system. Exploiting data streams to make a learning system is major discussion among such scenario [20] [21] [16] [22] [10] [23] [8] [24] [25] [26] [7] [17].

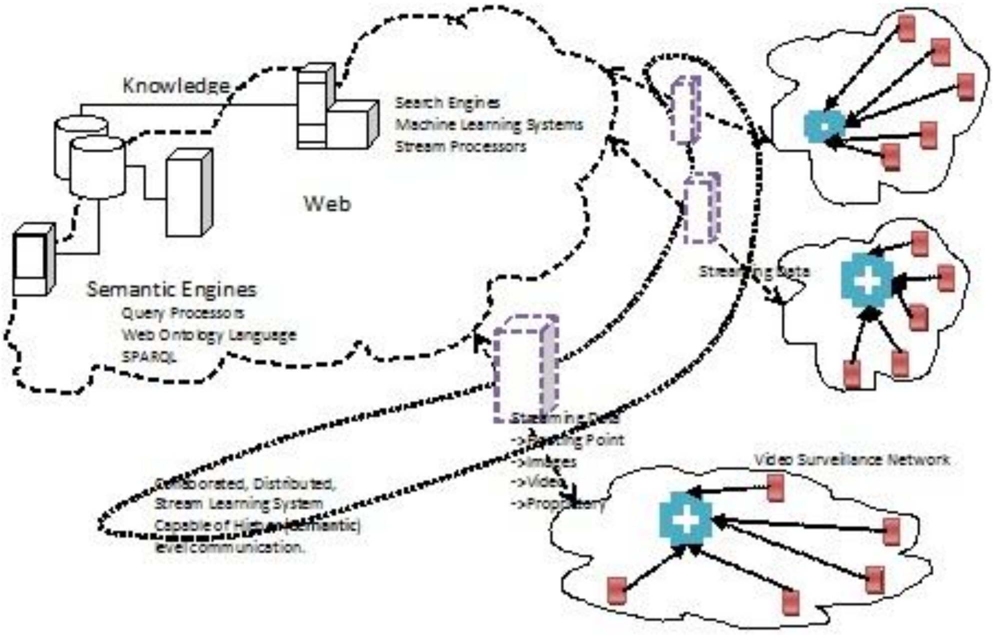

Figure 3 Scenario for Data Stream Learning in IoT

Ontology development for overall system has become inevitable [27]. Dimensionality reduction requires very powerful computers. This leaves out vast research area to implement such solutions at lower levels of data processing hierarchy [4]. Data abstraction challenge that will be faced by edge computing service discussed in [5] will remain prominent among related research areas in near future. Symbolic reasoning based on ontology is combined with complex event processing to develop a real-time stream reasoning approach [21]. Although in infancy, the approach is promising if further work in done in underlying methods for data extraction and abstraction. Extracting higher level information from raw sensory data has many applications [28]. Utilizing cloud resource has become reality now. Using cloud for analysis of IoT data [2] can give results but with latency that cannot be accepted in making quick decisions like evacuating some parts city based on rain water log and drainage failure.

Table 1 Literature Survey

No.	Area of Work	Related Word found in
1	Data Stream Learning	[24] [21] [6] [7] [15]
2	Problem Specific Implementation	[17][15]
3	Data Management Framework at Fog Layer	[13]
4	Survey on IoT with proposed framework	[13][24][1][9][2]
5	Semantic and Ontology Based Solution	[1] [21][29][28]

A tabular representation of how recent researches have focused on Data Stream mining is represented here.

4. CONCLUSION

This paper discusses state of research at various levels of network hierarchy for supporting an intelligent stream learning system. It introduces the basic concepts related to this research area. It then explores possibilities in existing research for desired model. It shows that for requirement specific applications, models have been implemented that apply Stream Learning concepts. Models can be developed based on specialized requirements in specific IoT environments. Machine learning techniques can be exploited to formulate mathematical models on stream data. The area of research for Data Stream Learning is emerging.

5. REFERENCES

[1] R. Ranjan, D. Thakker, A. Haller, and R. Buyya, “A note on exploration of iot generated big data using semantics,” 2017.

[2] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of things (iot): A vision, architectural elements, and future directions,” Future generation computer systems, vol. 29, no. 7, pp. 1645–1660, 2013.

[3] H. Cao, M. Wachowicz, C. Renso, and E. Carlini, “An edge-fog-cloud platform for anticipatory learning process designed for internet of mobile things,” arXiv preprint arXiv:1711.09745, 2017.

[4] F. Ganz, P. Barnaghi, and F. Carrez, “Automated semantic knowledge acquisition from sensor data,” IEEE Systems Journal, vol. 10, no. 3, pp. 1214–1225, 2016.

[5] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, “Edge computing: Vision and challenges,” IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637–646, 2016.

[6] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Wo´zniak, “Ensemble learning for data stream analysis: A survey,” Information Fusion, vol. 37, pp. 132–156, 2017.

[7] J. Demˇsar and Z. Bosni´c, “Detecting concept drift in data streams using model explanation,” Expert Systems with Applications, vol. 92, pp. 546–559, 2018.

[8] J. Sun, H. Zhang, A. Zhou, and Q. Zhang, “Learning from non-stationary stream data in multiobjective evolutionary algorithm,” arXiv preprint arXiv:1606.05169, 2016.

[9] C. C. Aggarwal, N. Ashish, and A. Sheth, “The internet of things: A survey from the data-centric perspective,” in Managing and mining sensor data. Springer, 2013, pp. 383–428.

[10] D. Anicic, S. Rudolph, P. Fodor, and N. Stojanovic, “Stream reasoning and complex event processing in etalis,” Semantic Web, vol. 3, no. 4, pp. 397–407, 2012.

[11] Y. Simmhan, “Big data and fog computing,” arXiv preprint arXiv:1712.09552, 2017.

[12] L. Atzori, A. Iera, and G. Morabito, “The internet of things: A survey,” Computer networks, vol. 54, no. 15, pp. 2787–2805, 2010.

[13] M. A. Abbasi, Z. A. Memon, J. Memon, T. Q. Syed, and R. Alshboul, “Addressing the future data management challenges in iot: A proposed framework,” INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, vol. 8, no. 5, pp. 197–207, 2017.

[14] M. Jaworski, P. Duda, and L. Rutkowski, “New splitting criteria for decision trees in stationary data streams,” IEEE transactions on neural networks and learning systems, 2017.

[15] D. Puschmann, P. Barnaghi, and R. Tafazolli, “Adaptive clustering for dynamic iot data streams,” IEEE Internet of Things Journal, vol. 4, no. 1, pp. 64–74, 2017.

[16] A. Haque, L. Khan, M. Baron, B. Thuraisingham, and C. Aggarwal, “Efficient handling of concept drift and concept evolution over stream data,” in Data Engineering (ICDE), 2016 IEEE 32nd International Conference on. IEEE, 2016, pp. 481–492.

[17] Y. Yamato, Y. Fukumoto, and H. Kumazaki, “Predictive maintenance platform with sound stream analysis in edges,” Journal of Information processing, vol. 25, pp. 317– 320, 2017.

[18] M. D. de Assuncao, A. da Silva Veith, and R. Buyya, “Distributed data stream processing and edge computing: A survey on resource elasticity and future directions,” Journal of Network and Computer Applications, vol. 103, pp. 1–17, 2018.

[19] R. Petrolo, V. Loscri, and N. Mitton, “Towards a smart city based on cloud of things, a survey on the smart city vision and paradigms,” Transactions on Emerging Telecommunications Technologies, vol. 28, no. 1, 2017.

[20] A. Haque, L. Khan, and M. Baron, “Semi supervised adaptive framework for classifying evolving data stream,” in Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2015, pp. 383–394.

[21] M. Endler, J.-P. Briot, F. S. E. Silva, V. P. de Almeida, and E. H. Haeusler, “An approach for real-time stream reasoning for the internet of things,” in Semantic Computing (ICSC), 2017 IEEE 11th International Conference on. IEEE,2017, pp. 348–353.

[22] M. Endler, J.-P. Briot, V. P. Almeida, F. S. E. Silva, and E. Haeusler, "Towards stream-based reasoning and machine learning for iot applications," 2017.

[23] E. Tu, N. Kasabov, and J. Yang, "Mapping temporal variables into the neucube for improved pattern recognition, predictive modeling, and understanding of stream data," IEEE transactions on neural networks and learning systems, vol. 28, no. 6, pp. 1305–1317, 2017.

[24] H. He, S. Chen, K. Li, and X. Xu, "Incremental learning from stream data," IEEE Transactions on Neural Networks, vol. 22, no. 12, pp. 1901–1914, 2011.

[25] P. P. Angelov, X. Gu, and J. Pr´ıncipe, "Autonomous learning multi-model systems from data streams," IEEE Transactions on Fuzzy Systems, 2017.

[26] Y. Sun, K. Tang, L. L. Minku, S. Wang, and X. Yao, "Online ensemble learning of data streams with gradually evolved classes," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 6, pp. 1532–1545, 2016.

[27] Y. Al-Hazmi and T. Magedanz, "Towards semantic monitoring data collection and representation in federated infrastructures," in Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on. IEEE, 2015, pp. 17–24.

[28] F. Ganz, D. Puschmann, P. Barnaghi, and F. Carrez, "A practical evaluation of information processing and abstraction techniques for the internet of things," IEEE Internet of Things journal, vol. 2, no. 4, pp. 340–354, 2015.

[29] J. Kiljander, A. D'elia, F. Morandi, P. Hyttinen, J. Takalo- Mattila, A. Ylisaukko-Oja, J.-P. Soininen, and T. S. Cinotti, "Semantic interoperability architecture for pervasive computing and internet of things," IEEE access, vol. 2, pp. 856–873, 2014.