# Information Abstraction and Knowledge Acquisition from Real-World Data

Frieder Ganz

Centre for Communication Systems Research, University of Surrey, GU17XH Guildford, United Kingdom
F.Ganz@Surrey.ac.uk

**Abstract**: This research focuses on enabling heterogeneous sensor devices to connect and communicate their data on the Internet and develops software systems than can interpret this data and create human understandable or machine-interpretable perceptions from the data. The software mechanisms collect and process the sensor data, identify different patterns and create higher-level abstractions from the data according to the context of the sensor device. This research develops some of the key components that are used as enabling solutions for the *Internet of Things*.

## 1 Main Objective and Motivation

There is a growing demand for new information processing techniques and methods that are able to process the myriad of data that is collected by sensory devices and mobile users. It is predicted that in the next 5-10 years there will be around 50 billion Internet connected devices that will produce 20% of non-video Internet traffic. The devices include sensor nodes, smart phones, GPS and many other sources that capture and communicate the real world data. This deluge of data requires efficient mechanisms to address the data communication and data management challenges for the future Internet.

My research introduces a new set of solutions for providing access to sensor data using a gateway component and creating multi-resolution representations of sensor data to decrease traffic in the real world data communication networks.

I also work on constructing patterns and abstractions that represent events and occurrences or other machine interpretable concepts and co-relations between the data items to create perceptions from the raw data. This perception can be used for situation-aware intelligent decision making and smart applications that use the real world intelligence to enhance their response to the changes and events in the physical world.

In this research, the data from physical world are captured and shared via common service interfaces that run on a gateway. This allows different users and data consumers to access the data (by following security and privacy procedures) and uses it directly or integrates it into their applications.

There is a growing interest in integrating this data into existing and next generation of smart applications that provide situation awareness and can make decisions or make recommendations according to the changes in the real world.

For example, a video projector that could turn itself on and off if people leave a conference room or a warehouse that could manage its stock automatically based on

communication with the items in the store, or monitoring the traffic patterns, people's movement, weather and other data in an urban environment and finding co-relation between different events that impact the traffic flow in different parts of a city.

This research addresses two major challenges:

- Developing an intelligent gateway to enables different sensory devices to provision and share data via Web services. This provides a homogeneous access to heterogeneous resources that facilitate sensing and capturing of the data from the physical world. The current gateway implementation provides plug-ins to support communication with the most common sensor node operating systems and platforms. It is also expandable to other platforms by developing and adding new plug-ins. The gateway also includes an automated association mechanism that enables sensor nodes to automatically connect and register themselves to a nearby gateway.

- Information analytics and creating human-understandable and machine-interpretable abstractions (i.e. perception) by using symbolic analysis and semantic processing of the observation and measurement data. The perception creation method interprets the observations and measurements by processing semantic descriptions of the data (i.e. type, location, time, quality attributes), analyses co-occurrences of different patterns from various relevant (i.e. from nearby location) and uses probabilistic machine learning techniques to identify relations between different data items and patterns in different time and locations. The result of the pattern analysis are then used in a logic inference mechanism (i.e. in the current research we use abductive logic) to create meaningful perceptions from the original data and the constructed patterns.

## 2 State-of-the-Art

Data aggregation and compression are the main solutions used to reduce communication traffic in IoT. In this section we review some of the common solutions and highlight the uniqueness of our data abstraction approach.

Chen et al. [1] create a summarized data stream of a set of sensory data streams and use the aggregated data for transmission. The summarization of the data relies on the mathematical sum, max, min, average and count aggregate functions [2].

A different approach to reduce the communication traffic in communicating the sensory data is to reduce the size of the messages. This can be realized using data compression algorithms. However, compressing the data itself could lead to a loss of information (in lossy compression) and the compression techniques can require higher power consumption as compression requires data processing before transmission and in long term observations (e.g. environmental monitoring applications) compression techniques data can still create large amounts of data [3].

Wang et al. [4] use compression techniques with adaptive sensing for WSN. This can be exploited to transmit only very dense (and therefore small in size) data and then reconstruct the overall "data" by applying a reverse function. However, the constructed data relies on several incomplete data-streams and therefore this can lead to a huge loss in the quality of the reconstructed information.

A new way to reduce data communication in Real World data networks is to extract the information which is relevant or important for the user before transmitting it. However, determining what is required or what is important to the user from heterogeneous data sources and in the wide range of applications that IoT and Cyber-Physical systems can use is not a trivial task. So it is important to define methods that can create higher-level abstractions that can be general purpose and then to use abstract reasoning models that can transform these abstractions into machine interpretable or human understandable knowledge.

Yun et al. [5] introduce a similar approach that exchanges information using ”signatures” instead of the raw data. The signatures are combinations of properties measured during an event such as ”bright light” and ”loud sound” during the explosion of a bomb. The signatures consist of a string representation indicating that something is present or absent at a particular sensor. This is efficient in terms of data communication as only binary data has to be transmitted such as ”NYY” standing for ”No” - light is not present and ”Yes” sound and temperature are present; however is not precise in terms of defining various patterns of data and its extendibility and scalability is also limited.

In our approach, we discretize the data and create pattern representations of the source data which can be used to describe the transient states of a sensor data stream (e.g. low noise (A), medium noise (B), high noise (C) resulting in ABC instead of binary No). Then these patterns are fed into a probabilistic reasoning model to create higher-level abstractions from the data that is emerging from the sensor data streams.

## 3 An Information Abstraction Framework

We have implemented our solution on a gateway component developed in our previous work [6]. The gateway component provides a connection between heterogeneous sensor devices with low processing and communication capabilities and higher-level services and applications on the Internet. The gateway can be equipped with several air interfaces such as IEEE 802.15.4, IEEE 802.11 and Bluetooth to support a variety of communication links over the physical layer. The network protocols such as 6LoWPAN [7] and Zigbee protocol stack [8] are also supported to enable network layer communications. The data management in the gateway is divided into three main tiers (shown in Figure 1) that include data collection, data processing and data provisioning.

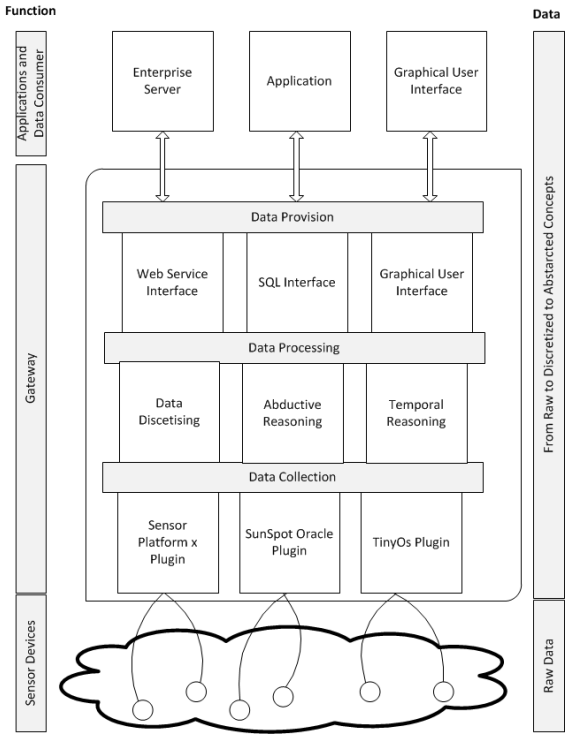


**Fig. 1.** A layered architecture for information abstraction system

### 3.1 Data Collection

The data collection tier provides wrappers for different hardware and software platforms and supports communication between different sensory devices and the gateway. The wrappers are implemented as plug-ins for different data sources. The data collection layer provides an abstract view and hides the complexity of underlying devices (i.e. sources) and the proprietary hardware and software specifications.

The wrappers include a protocol for negotiation and association of sensory devices to the gateway [6]. The modular design of the system makes it easier to automatically setup and connect a large amount of heterogeneous sensor sources. In the current gateway different plug-ins are implemented for TinyOs, Contiki enabled devices and Oracle SunSpot nodes.

### 3.2 Data Processing

The data processing tier uses caching and storage functions to minimise direct interaction with the sensory devices. The important aspect of the gateway is the ability to process and interpret the data and create higher-level abstractions. This provides a local computing and data abstraction and a global data/concept communication para-

digm which allows more efficient communication and integration of large sensory data.

The data processing tier consists of three sub components: data discretizing, abductive reasoning and temporal reasoning. The data discretizing component is used to discretize the raw sensor data into lower-dimensional representations. This component utilizes an extended version of the Symbolic Aggregate approXimation (SAX) algorithm [9], called SensorSAX optimised for sensor data, to convert continuous data (e.g. f1,2,3,4,5,4,3,2,1g) into a compressed discretized representation (e.g. fa,b,b,ag).

The abductive model stores the mapping between discretized representation and abstractions (e.g fa,b,c,dg represents ”attendance” in a room). The abductive reasoning and temporal component infers the current observations and determines which abstractions are the most plausible ones.

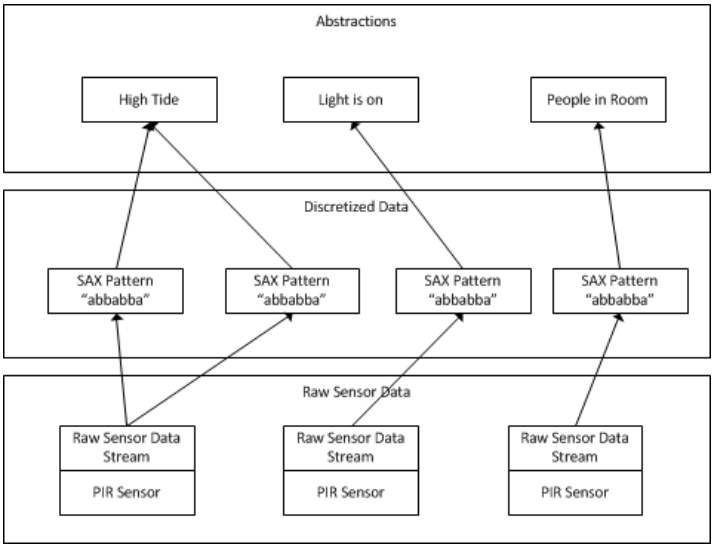


**Fig. 2.** Processing from raw data into abstraction

### 3.3 Data Provision

The data provision tier provides different interfaces for the data access such as Web service interfaces, and APIs to query and retrieve the abstracted concepts for traffic efficient communications. It also provides direct raw data access if it is required.

## 4 Evaluation

To prove the feasibility of the proposed approach, we measure accuracy, data size reduction and latency of the process to create the abstractions. The gateway collects the data from several stations and provides the abstracted information. We evaluate

the accuracy of the method by comparing the constructed abstractions.. We measure the data size reduction (i.e. we only measure the data size and other communication protocol overheads are not included) and calculate the average correlation between the original and reconstructed data. Current results show that the data has a correlation coefficient of 0.89 with a positive direction which means that the reconstructed data is very similar to the original data. The execution time is also measured for the construction of a set of abstractions over a data collection period.

## 5 Outlook

The future work will focus on the parameter learning for the probabilistic model by applying expectation maximization (EM) algorithms to increase the accuracy of abstractions.

## 6 References


1. Y. Chen, J. Shu, S. Zhang, L. Liu, and L. Sun, "Data fusion in wireless sensor networks," in Electronic Commerce and Security, 2009. ISECS '09. Second International Symposium on, vol. 2, may 2009, pp. 504 –509.
2. P. Jesus, C. Baquero, and P. S. Almeida, "A survey of distributed data aggregation algorithms," The Computing Research Repository ACM, vol. abs/1110.0725, 2011.
3. N. Kimura and S. Latifi, "A survey on data compression in wireless sensor networks," in Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on, vol. 2, april 2005, pp. 8 – 13 Vol. 2.
4. J. Wang, S. Tang, B. Yin, and X.-Y. Li, "Data gathering in wireless sensor networks through intelligent compressive sensing," in INFOCOM, 2012 Proceedings IEEE, march 2012, pp. 603 –611.
5. M. Yun, D. Bragg, A. Arora, and H.-A. Choi, "Battle event detection using sensor networks and distributed query processing," in Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on, april 2011, pp. 750 –755.
6. F. Ganz, P. Barnaghi, C. Francois, and K. Moessner, "Context-aware management for sensor networks," in the Fifth International Conference on COMmunication System softWAre and middlewaRE (COMSWARE11), ACM, 2011.
7. G. Mulligan, "The 6LoWPAN architecture," in Proceedings of the 4th workshop on Embedded networked sensors, 2007, p. 7882.
8. ZigBee, "ZigBee specifications," 2010. [Online]. Available: http://www.zigbee.org/Specifications.aspx
9. J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, ser. DMKD '03. New York, NY, USA: ACM, 2003, pp. 2–11.