Mining sensor data from complex systems

Vespier, U.

Citation

Vespier, U. (2015, December 15). *Mining sensor data from complex systems*. Retrieved from https://hdl.handle.net/1887/37027

Author: Vespier, Ugo
Title: Mining sensor data from complex systems
Issue Date: 2015-12-15

Chapter 2

Sensor Data and Complex Systems

In this chapter, we discuss the context of this dissertation and the motivation behind the presented work. We also present the InfraWatch project, the main application and testbed for the methods discussed in the next chapters.

2.1 Big Data

As mentioned, the term *Big Data* has received a lot of attention over the last years [64], although it is often cause of confusion as its meaning is not always well-defined in all contexts. Big data is a broad term that refers to the challenges in managing and analyzing large quantities of data.

A widely accepted definition of Big Data has been given by Gartner's analysts Beyer and Laney in a 2012 industry report [11] where the authors define the term in relation to three main challenges:

Volume The ever-growing amount of data that institutions and companies have to deal with offers serious challenges in terms of data management and storage. Real-world examples range from the data collected by astronomers with radio telescopes, to the massive amount of messages exchanged nowadays on social networking platforms such as Facebook and Twitter. Novel storage and indexing methods, able to cope with this huge amount of collected information, need to be employed.

Velocity A second important challenge is related to how long it takes to process incoming data. For example, a credit card institution analyzing massive streams of incoming transactions would like to detect potential fraud without delay. Real-time processing methods are needed in order to cope with the *velocity* challenge.

Variety Last but not least, Big Data comes in any type, both in structured and unstructured form. Text, audio, video, sensor data, log files, and combinations of these, are examples of data types found in Big Data applications [15]. The development of methods able to cope with this broad variety of data is another challenge posed by Big Data applications.

A great extent of the efforts aimed at solving big data problems revolve around these three challenges, as well as a broad range of applications across several science fields and industries. Sensor networks and monitoring is an important one and represents the main focus of this thesis. In particular, the rise of the Internet of Things [4, 5] is directly connected to the challenges posed by the management and analysis of Big Data.

2.2 Sensor Networks and the Internet of Things

The last two decades have witnessed a tremendous growth in the availability of sensor data collected from a multitude of systems in various application domains [3, 16]. In fact, the wider availability of cheap sensor technology has enabled large sensor networks that continuously monitor and analyze physical systems such as infrastructures, cars [25, 45, 59, 32], airplanes [7, 107, 9, 87] and, last but not least, the human body [14, 83].

The increasing presence of sensing devices coupled with the pervasiveness of connectivity is paving the way for a scenario in which physical objects, humans and software communicate to achieve common goals by directly interacting with the physical world. This paradigm is called Internet of Things (IoT) [35, 5] and several applications of its concepts are already used in production.

In a recent report [72] by McKinsey & Company, the authors categorize IoT applications in three main areas:

- Tracking behavior
- Enhanced situational awareness
- Sensor-driven decision analytics

Companies and institutions are interested, first of all, in tracking and monitoring products and objects in real-time [54] in order to increase the efficiency of their operations or fine-tune their business or pricing models. Consider, for example, the case of a car insurance company that installs sensors in their customers' cars in order to monitor and model the behavior of drivers [90].

Situational awareness [106] is another key application area of the IoT. Large sensors networks, in fact, can be deployed in infrastructures, such as roads, buildings or bridges, or installed in certain areas to report on environmental conditions. Data coming from the sensors can then be used to enhance the awareness of decision makers about the observed events in real-time, especially when data is coupled with tailored visualization technologies.

Ultimately, sensor networks can support long-term and complex decision making. In the retail industry, for example, some companies are experimenting with sensors that continuously monitor shoppers [67, 63] as they move through stores in order to measure how long they stand in front of any given display and correlate it with what they ultimately buy. Data of this kind can help, in the long run, by optimizing store layouts and increase revenues.

Especially when long-term decision support is of interest, sensors networks produce large volumes of data continuously over time. This opens up several challenges both from a data management and from a data analysis perspective. Modern sensor networks have to be supported with state-of-the-art data management solutions in order to cope with the large amount of collected data and ensure its effective storage, access and visualization.

Such large amounts of data, on the other hand, represent an opportunity to apply data mining methods to better understand the observed system and get insight into its behavior [1]. Moreover, as these data sources continuously provide data over time, the research community is also focusing on methods and algorithms to analyze information in a streaming fashion in order to provide real-time insights [31].

2.3 Multi-Scale nature of Complex Systems

Sensor networks are often employed to monitor and analyze complex, dynamic systems, which exhibit non-obvious behavior. An important example are systems affected by several phenomena at different temporal scales.

Consider, for example, the electrical system of an apartment whose aggregate power consumption is measured by a smart meter [108, 58]. The time series data collected by the smart meter would be affected by all the operational home appliances, heating units and lighting systems in the house. As different home appliances are switched on at different times and have diverse operational durations, their effect on the aggregate consumption can range from short-term spikes, for example in the case of a boiler, to longer, more equally distributed patterns, as for example in the case of a washing machine [73]. For example, the time series in Figure 2.1 shows four days of power usage from an apartment in the Smart* dataset [8]. In the data, there is a clear long-term periodic component due to the cycles of the refrigerator. Shorter-term patterns, however, show up superimposed in the data and correspond to activations and deactivations of the various electrical appliances in the house.

Another real-world example borrowed from civil engineering regards dikes [20, 70]. In dikes, it is typically of interest for civil engineers to measure the water pressure at specified location of the infrastructure. The amount of pressure depends on several factors and, also in this case, it is possible to classify them

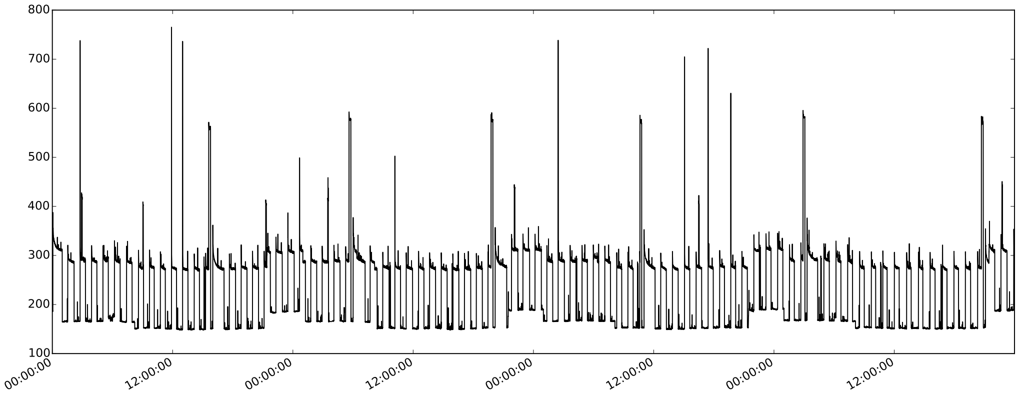

Figure 2.1: Four full days of power usage (in watts) from one of the houses in the Smart* dataset [8].

along diverse temporal scales. For example, lunar tides indirectly affects water pressure following both a half-daily cycle and a longer-term, two-weekly cycle.

Analysing systems such as the ones described above requires methods capable of dealing with the presence of multiple relevant scales of analysis in order to extract insights at all levels and resolutions. This represents one of the main challenges addressed in this thesis.

2.4 SHM and InfraWatch

One relatively recent application of sensor networks and sensor data analysis is the monitoring of infrastructural assets such as bridges, tunnels, etc. [21]. In fact, according to a recent survey from the US Federal Highway Commission [28], on average 56% of the assessments to civil infrastructures made by visual inspection are inappropriate, suggesting additional methods of monitoring to guarantee the safety of the assets.

Structural Health Monitoring (SHM) is an interdisciplinary field at the intersection between civil engineering, signal processing, sensor technology, material sciences, data management and mining, which is emerging in order to find alternative or complementary solutions to the visual inspection. In fact, the use of advanced sensing and monitoring systems provides the opportunity to collect real-time information from infrastructures, in order to monitor their performance and to deduce relevant knowledge for decisions on their maintenance demand [26, 86]. Asset owners can use this information to assess the life time perspective of (crucial) infrastructural links and to plan the window within which maintenance can be conducted. When considering the stock of infrastructural assets in view of service-life assessment, monitoring and sensing systems are very valuable instruments that can be used to extract actual information about its condition and performance.

In typical SHM scenarios, sensor systems are mounted in or to structures and monitor the environmental as well as the internal condition of the measured system over long periods of time. The collected sensor data is typically continuously analyzed in order to detect inconsistencies or anomalies in how the structure is behaving and notify potential problems in time. Aside from notifying anomalies, SHM systems are also used to monitor and forecast degradation mechanisms in order to plan maintenance in a more informed way.

In the next section, we present a particular SHM project in detail. This project and its data will serve as a testbed for a great extent of the methods and algorithms presented in this thesis.

2.4.1 The InfraWatch project

InfraWatch is a project that is part of a Dutch STW's funded program called Integral Solutions for Sustainable Construction (IS2C). The program is composed of nine research projects with the common goal of setting new standards and advancing the state of the art in the field of sustainable construction and service-life assessment.

As part of the IS2C program, InfraWatch[1] focuses on sensing, monitoring and degradation mechanism from a data analysis perspective. Subject of the project is an important Dutch highway bridge: the Hollandse Brug. The Hol-

---

[1] http://www.infrawatch.com

Figure 2.2: Picture of the Hollandse Brug, which connects the ‘island’ Flevoland to the province Noord-Holland.

landse Brug is a bridge between the Flevoland and Noord-Holland provinces and is located at the place where the Gooimeer joins the IJmeer (see Figure 2.2). The bridge was opened in June 1969 and National Road A6 uses it. There is also a rail connection parallel to the highway bridge, as well as a lane for cyclists on the west side of the car bridge.

In April 2007, it was announced that measurements would have shown that the bridge did not meet the quality and security requirements. Therefore, the bridge was closed in both directions for heavy traffic on April 27, 2007. The repairs were launched in August 2007 and a consortium of companies, Strukton, RWS and Reef has installed a monitoring configuration underneath the first south span of the Hollandse Brug with the main aim to collect data for evaluating how the bridge responds to load. The sensor network is part of the strengthening project which was necessary to upgrade the bridge’s capacity by overlaying.

The monitoring system comprises 145 sensors that measure different aspects of the condition of the bridge, at several locations along the bridge (see Figure 2.3 for an illustration). The following types of sensors are employed [55, 56]:

- 34 'geo-phones' (vibration sensors) that measure the vertical movement of the bottom of the road-deck as well as the supporting columns.
- 16 strain gauges embedded in the concrete, measuring horizontal longitudinal strain, and an additional 34 gauges attached to the outside.
- 28 strain gauges embedded in the concrete, measuring horizontal strain perpendicular to the first 16 strain gauges, and an additional 13 gauges attached to the outside.
- 10 thermometers embedded in the concrete, and 10 attached on the outside.

Furthermore, there is a weather station, and a video-camera provides a continuous video stream of the actual traffic on the bridge. Additionally, there are also plans to monitor the adjacent railway bridge.

The current monitoring set-up is clearly providing many challenges for data management. The 145 sensors are in fact producing data at rates of 100 Hz, which can amount to a gigabyte of data per day. Adding to that is the continuous stream of video.

Project goals and expectations

InfraWatch is, primarily, a Structural Health Monitoring project and its goals are directly related to questions about the observed infrastructure from a civil engineering perspective. The following tasks, in particular, are of importance to the civil engineers involved in the project:

- obtain a summary of the major phenomena affecting the bridge infrastructure over time and their impact.
- given historical sensor data from the bridge, obtain a qualitative and

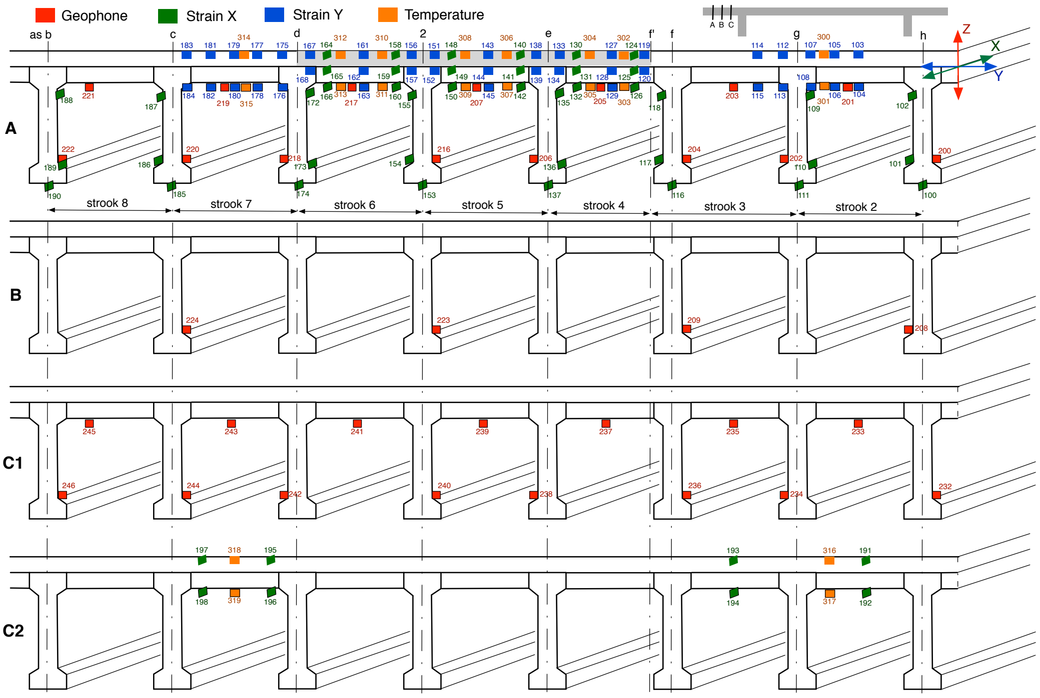

Figure 2.3: Diagram explaining the individual sensor placement on the Hollandse Brug.

quantitative estimate of the structural health of the Hollandse Brug.

- given the current sensor network deployment, obtain a new configuration of sensors, possibly employing fewer sensors, that is equivalent or comparable in terms of collected information.

In this thesis, we provide fundamental data mining methods and algorithms that can be employed to design a solution for the tasks above and related tasks involving sensors monitoring and analysis of systems.

Technical challenges

In order to make the goal of InfraWatch feasible from a data mining perspective, we need to identify a set of technical tasks which could serve, paired with domain knowledge in civil engineering, as basic tools to provide a solution. Moreover, as the bridge is affected by phenomena of different nature at different temporal scales, the methods and solutions will have to cope with its multi-scale nature. The tasks we are interested in are described below:

- given a sensor-based time series, identify which are the relevant temporal scales of analysis.
- given a sensor-based time series, identify which are the recurring patterns in the data at multiple temporal scales.

In addition to the tasks above, we are also interested in visualizing effectively the large amount of time series data produced by the InfraWatch project.