First Middle East Conference on Smart Monitoring,
Assessment and Rehabilitation of Civil Structures
8 – 10 February 2011, Dubai, UAE

SMAR 2011

Large Data Stream Processing for Bridge Management Systems

Arno Knobbe[1], Arne Koopman[1], Joost Kok[1], Bas Obladen[2], Carlos Bosma[2] and Eddy Koenders[3]

[1] LIACS, Leiden University, the Netherlands
[2] Strukton, the Netherlands
[3] Microlab, Delft University of Technology, the Netherlands

ABSTRACT: This paper introduces a new research project called InfraWatch[2] that demonstrates the many challenges that a large complex data analysis application has to offer in terms of data management and analysis. The project is concerned with intelligent monitoring and analysis of large data streams derived from large infrastructural projects in the public domain, with the specific aim of Structural Health Monitoring. The project focuses on an important highway bridge in the Netherlands, which is currently equipped with a multitude of vibration and strain sensors, a video camera and weather station. Within the framework of this paper, many large scale data mining settings will be considered, such as analysing patterns in the streams of sensor output, the discovery of relations between different sensors, as well as analyzing trends over time. The paper will provide an overview of the large data streams that are involved and how relevant management information can be distilled.

1 INTRODUCTION

In this paper, we consider the challenge of dealing with the large volumes of data generated by sensor systems installed on large infrastructural assets. With a variety of sensors becoming ever cheaper, and the cost of data storage and processing decreasing, it is quickly becoming attractive to fit both new and existing bridges, tunnels and so forth, with large collections of sensors that continuously monitor the physical and structural state of various details of the infrastructure. Because sensors often have the potential of measuring at high frequency, an asset fitted with several hundreds of sensors can produce considerable streams of data around the clock, and serious attention will need to be given to storing and transporting this data. In this text, we sketch some possibilities for how to manage this deluge of data, as well as how this potentially rich source of data can be exploited for monitoring the health of the infrastructure and managing its use by the public.

We assume here that we are dealing with an asset that is fitted with a generic sensor network. That is, the network is not so much designed for one specific application, but rather should be able to support a range of goals, with different requirements on the number of sensors involved, the measurement interval and so in. This means that the data management strategy adopted cannot assume any limitations in scope, and should thus be designed to handle the maximum volume of data coming in from the sensor collection. In fact, the specific implementation that inspired our work (of which more below) assumes that *all* data produced by the sensors at the

[2] InfraWatch is part of the national STW Perspectief program "Integral Solutions for Sustainable Construction" (IS2C).

highest reasonable frequency is being handled and stored off-site, to allow for every conceivable future analysis question. Of course, a number of applications that rely on this sensor network require only a subset of this data, but often, these applications will need to work alongside each other, such that the limitation of scope of each individual application does not necessarily reduce the total volume of data.

As an example of how different applications may pose different requirements on the scope of the data, consider an application that examines seasonal effects in traffic over a bridge, as well as a monitoring application that is allowing human managers to get a high-level impression of the recent state of the bridge, possibly including a live video stream of the traffic over the bridge. For the monitoring application, no long term storage is required, and only a small selection of sensors may be required, producing fairly low-frequency data. Although this application deals with only reasonably small volumes of data, it does however require the data to be very recent, if not instantaneous. Therefore, transporting data periodically from the site to the monitoring office is not acceptable. In the seasonal analysis application, on the other hand, it may be quite acceptable to not have the latest data, as long as data over a long period of time (including multiple seasons) is available, and all sensors are included in the data. Because the long-term analysis should not get in the way of the on-line monitoring, the data management should be flexible enough to allow multiple applications with different data scopes.

To characterize the volume of data coming from a sensor network, or analogously, the data scope of an application, one can essentially use the following four variables:

- *Frequency*. A higher frequency obviously leads to a larger data volume.
- *Number of sensors*. As mentioned, we assume a network comprised of many, typically cheap, sensors, such that a range of applications is supported. The sensors will often exhibit certain levels of redundancy
- *Measuring interval*. We assume the system to work over the entire lifespan of the asset or the sensors, although specific application will typically not involve all this data.
- *Latency*. Shorter latencies allow for real-time operations.

In this paper, we do not limit the discussion to the challenges of large data stream management. We also sketch a number of both well-established as well as novel data analysis techniques to support a number of information needs owners of the asset may have. The techniques vary in their computational complexity, but even for the simplest analysis techniques, special computational facilities will be required in order to operate them on the gigabytes of data that can be expected from networks as described above.

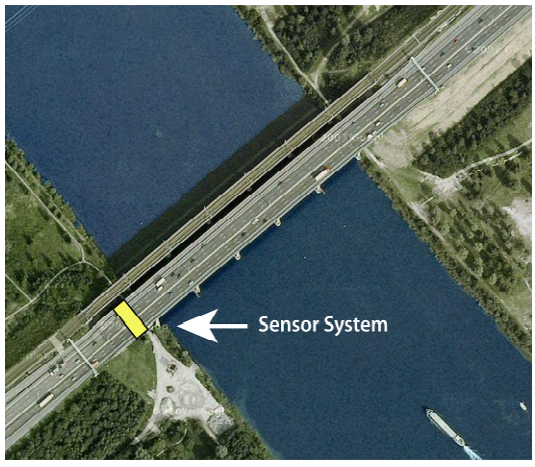

Figure 1. Aerial picture of the situation of the Hollandse Brug, which connects the island Flevoland to the province Noord-Holland.

1.1 The InfraWatch Project

The InfraWatch project, which inspired this work, is centred on an important highway bridge that is already producing substantial quantities of data: the Hollandse Brug. This bridge is located between the provinces Flevoland and Noord-Holland in the Netherlands, which is where the

Gooimeer joins the IJmeer (see Figure 1). The bridge was opened in June 1969, and since then is used by national Road A6. There is also a connection for rail 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 to freight 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 intent of obtaining a state-of-the-art Structural Health Monitoring system. This sensor network is part of the strengthening project which was necessary to upgrade the bridges' capacity by overlaying.

The monitoring system comprises 145 sensors that measure different aspects of the condition of the bridge, at several locations on the bridge. The following types of sensors are employed:

- 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, which provides a continuous video stream of the actual traffic on the bridge.

Clearly, the current monitoring set-up is already providing many challenges for data management. For one, the 145 sensors are producing data at rates of 100 Hz, which can amount to a few gigabytes of data per day. This does not include the data that represents the continuous stream of video. Although the InfraWatch project is in its early stage, data is already being gathered and monitored. However, the current data available for analysis consists of short snapshots of strain and video data, which is being manually transported from the site to the monitoring location (typically an office environment or Leiden University).

2 DATA MANAGEMENT

As is clear from the introduction and the description of the Hollandse Brug situation, sensor systems for bridge monitoring can produce substantial quantities of data. Unfortunately, the amount of data produced at the Hollandse Brug does not allow the storage of data on-site for any long period, which makes data management a vital component. Furthermore, for on-line monitoring of the bridge, a low latency for at least some of the data is required. The generic data management configuration we propose for structural asset monitoring, which is also the intended end-situation at the Hollandse Brug, is depicted in Figure 2. This diagram clearly shows the main bottleneck in the system, which is the low-bandwidth connection between the on-site sensor system and the servers that take care of storage and off-line analysis.

Before discussing the details of the diagram, let's first consider the typical volume of data produced at the Hollandse Brug. The 145 sensors measuring at 100 Hz produce around 56 kB of data per second. This amounts to about 5 GB per day, and over 1.7 TB on a yearly basis. This number is based on a fairly efficient data representation, whereas in a straightforward text file

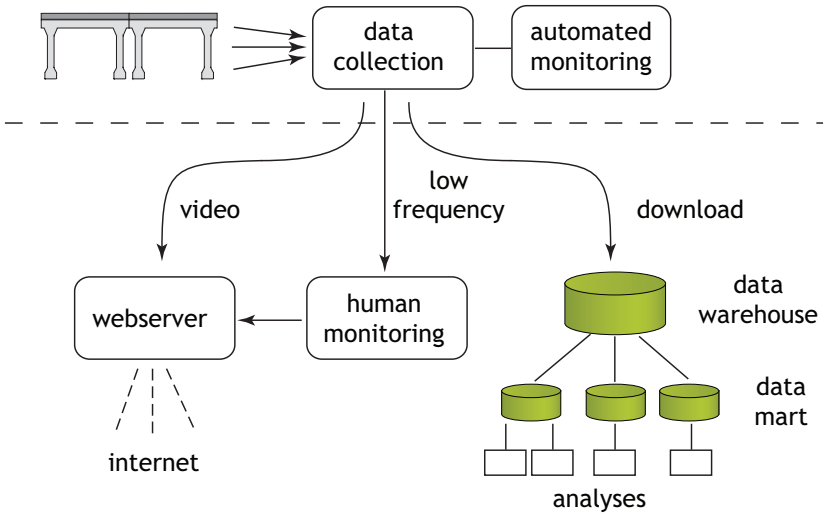

Figure 2. The generic data management configuration

representation, this would be at least three times as much. The video camera produces a data stream in a similar range, with 46 kB/s of compressed video, for a typical daytime situation. The diagram in Figure 2 shows both video and sensor data coming over the same connection. However, the nature of these two streams is quite different. Video will typically be used with very low latency, in order to show the current state of traffic flow. Also, video at high frame rates, and thus high quality, will only be used sporadically, and on specific request of the bridge owners. Apart from this occasional use of bandwidth, a live stream of the video will be made available through the project website at lower frame rates. This video stream will be accompanied by a low-frequency data stream of selected sensors, both for monitoring purposes and publishing to the Internet. Neither low-frequency stream will be stored off-site for longer duration.

The data streams described so far will result in a moderate but constant use of bandwidth, with occasional periods of intense use for high-quality video. The remaining bandwidth will be available for downloading the high-frequency sensor data to the *data warehouse* for permanent storage. As immediate and constant access to this data off-line is not required, the downloading of recent intervals of data can be scheduled, and any delays due to monitoring can be resolved by downloading in more quiet times of the day (typically at night). The complete sensor data is stored in a central data warehouse, but a typical analysis, using the techniques described in the next section, will often take place on specific selections of the data, so-called *data marts*. These data marts may for example concern data for a selected period, a specific selection of sensors, or data that has been down-sampled. Different analysis techniques may also require different representations. The data storage component supports all these alternative versions of the original data.

Finally, Figure 2 shows an automated monitoring component that is designed to analyse data instantaneously for detecting previously defined events and patterns. The specific nature of the events to be monitored can either be defined by the end-user, or can be derived by off-line analysis procedures. These procedures analyse specific selections of historic data in order to determine for example what constitutes an unusual event. The results of this analysis are then occasionally uploaded to the monitoring component. This component uses the handcrafted or derived definitions to monitor the occurrence of events, and either logs and counts them, or sends an alarm to the bridge owner.

3 DATA ANALYSIS

Although the management of huge quantities of data is a challenge in its own right, its analysis poses an even greater challenge. Many traditional data mining algorithm implementations, both research and industrial, are typically not suited to analyse terabytes of (stream) data. A straightforward approach that moulds the data to suit these implementations, by either cropping or sampling, would lose many of its interesting regularities. For example, if we sample the data we would preserve the effects on a monthly basis, but lose insight in how the infrastructure behaves dynamically on a single heavy-weight vehicle crossing the bridge. Ideally, the long- and short-term effects should both be available for analysis. Therefore, the aim for the InfraWatch project is to transform existing techniques and develop new techniques that can operate directly on terabytes of data, and allow for analysis on different time scales.

Even for small fragments of data, many data mining algorithms would not operate directly on high-frequency continuous data streams. Typically, one would pre-process the data in order to discriminate different characteristic events. For example, one could discretise the measurement data in the amplitude domain by using thresholds, such that heavy weight trucks get assigned a distinct value (see Figure 3 top). While this approach works for fixed-sized databases, long-term effects could lead to slow drifts in our large streams, which can as a result lead to errors in our resulting data after pre-processing (see Figure 3 bottom).

3.1 Analysis Tasks

In data mining, there are many types of analysis tasks that are studied, even when a single dataset or application is given. In this respect, the InfraWatch project is no different. In this section, we will outline some of the analysis tasks that we deem relevant in its context.

One well-known type of data mining task is that of mining *patterns* in data [Aggarwal]. Pattern mining focuses on finding regularities in the data that typically exhibits some short-term behaviour. As an example, if we consider our measurement data, one of the many patterns could be the reoccurring and similar sensor data that is recorded when a vehicle crosses the bridge. While pattern mining on large fixed-sized databases already is faced with some challenges, its application to streams is even harder. Since with streams, one typically does not have access to all data at the same time to see whether a pattern occurs regularly in the data or not.

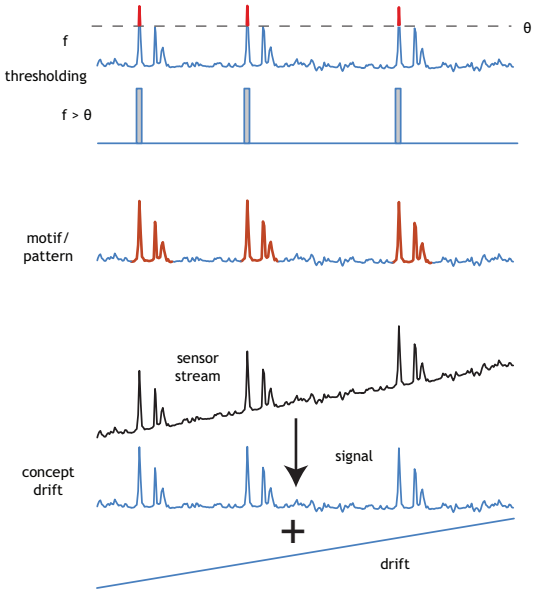

Figure 3. Various analysis methods that can be applied on time series data such as derived from the Hollandse Brug.

In contrast, the availability of recorded data over long periods of time, such as in the InfraWatch case, allows for the discovery of long-term effects. The high frequency signals which harbour the short-term patterns can be often superimposed on low frequency signals. For example, the average signal values can drift over time to a higher average. Concept drift often results in the degradation of the performance of an earlier well-performing model of short-term events [Žliobaite, 2010]. In order to adjust the short-term model, we therefore need to identify the drift

automatically. However, the drift itself can be a very interesting signal to analyse as well. In our bridge example, a drift of the characteristic frequencies can be an indication whether the structure as a whole is beginning to deteriorate.

Yet another analysis task focuses on the number of sensors that are installed on the bridge. Most streaming applications utilise only a single stream, which contrasts greatly to our 145 continuous concurrent streams. Moreover, we have additional background knowledge that we can utilise in our analysis task. On the Hollandse Brug, we have installed several types of sensors at specific locations, and each of them measures a specific physical property of the system (that is, strain, vibration, etc.). Each sensor type therefore provides different insight in the system, and the usefulness of each of these has to be analysed. Within the project, we will develop new sensor types with the specific aim to increase the efficiency of the models that can be learned from the system.

Building on this, we aim to develop a minimalistic system that performs well. That is, can we select a small set of sensors that leads to models that can do equally well in the prediction of the infrastructure's degradation. A minimalistic system would be ideal to deploy at future infrastructure sites that could undergo a similar form of monitoring. On the Hollandse Brug, we have deliberately installed an abundance of sensors, in order to allow for a range of applications, but this could be too costly for other settings. In the early stages of the InfraWatch project, we have already focussed on finding a small set of relevant installed sensors, which can serve as 'prototypes' for monitoring purposes. Based on a small data sample, we showed that some sets of sensors demonstrated similar events, which is an indication that we could reduce the number of sensors effectively [Koopman et. al., 2010].

Since 2008, the team at Strukton has been gathering a huge collection of data at a high frequency. This vast dataset allows for the discovery of many different effects on a long or short term. In the context of traffic modelling, it is interesting to see the effects of traffic jams, which typically occur at specific time periods on working days. While these patterns would occur on a daily basis, weekly traffic trends would also be interesting to see. For example, a holiday would typically stand out given a weekly pattern in the data. Moreover, seasonal effects and even effects that span over multiple years can already potentially be discovered.

3.2 Analysis Methods

Patterns or *motifs* are regularly occurring subsequences within a time series as for example demonstrated in Figure 3. These patterns can be of use for a large number of applications; in the InfraWatch case, the availability of different patterns over time in the stream can be an indication that the infrastructure is undergoing some type of change. In order to successfully apply pattern mining to the InfraWatch case, we need to focus on algorithms that can handle large scale databases.

A successful approach to motif finding is based on the SAX-encoding of continuous time series. This encoding reduces the dimensionality of the sequence both in amplitude and time. Based on this, [Patel et. al., 2002] have developed a method to efficiently find similar motifs within a large time series. This starting point has lead to many different algorithms that can deal with various cases, such as on-line discovery and exact or approximate motifs [Mueen et. al., 2009, 2010]. Within the InfraWatch setting, both on-line as off-line methods need to be extended and developed such that they can deal with the multi-dimensionality of the data, that is, the number of concurrent sensors.

Engineers typically have access to a large toolbox of building blocks and rules of thumb that form the basis for their design. However, in order for these to be applicable to a specific

situation one needs to choose the proper parameters. Consider that we have a system, with sensors *s*, which are related to each other through their position on the structure. Since the sensors are physically linked with each other, it is likely that this influence will be visible in the measurements of some of these sensors.

Say that an engineer would assume that there are desired relations between these sensors in the form of the following equation:

$$f_x(t) = c_0 + \sum_{s_y \in S} c_y \cdot s_y(t).$$

If the goal would be to select a type of concrete such that we would have specific relations between sensor measurements, the engineer would focus on fine tuning specific values of *c*. However, for an existing bridge with sensors installed, we need to do exactly the opposite. What can be seen as a form of reverse engineering, given our sensor system, we can ask the question: *how are the sensors related*? Instead of engineering the values *c*, we need to derive the values of *c* given the sensor data.

Based on an initial sample of the Hollandse Brug measurement data, we aimed to find relations in the form of equations between the installed sensors. In order to derive such equations, we used the Lagramge system to fit equations on our sensor data [Todorovski and Dzeroski 1997]. Such a procedure typically results in many equations that can be fitted on the data, all of varying quality. Using a greedy selection procedure, we selected a qualitative and small set of equations that forms a describing model for the sensor system. Although this can be considered as a small experiment in light of the huge volumes of data, already some interesting equations became visible [Koopman et al., 2010].

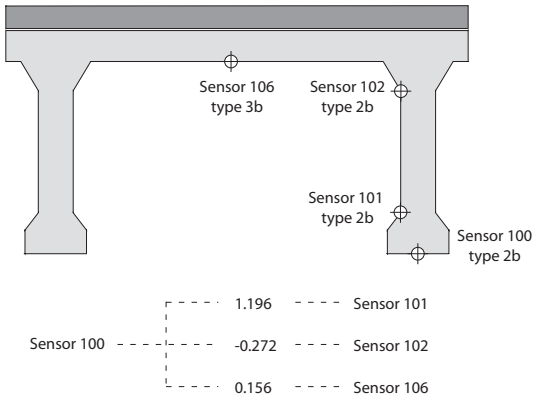

Figure 4. Shown is a detail of the cross section of the Hollandse Brug with some of its sensors (top). One sensor is strongly correlated with a set of other sensors in the system (bottom).

As shown in Figure 4, we see an example of such a relation between a set of sensors that is described by the following equation:

$$s_{100}(t) = 1.196 \cdot s_{101}(t) - 0.272 \cdot s_{102}(t) + 0.156 \cdot s_{106}(t) + 23.30.$$

Our procedure, which at present does not yet incorporate background information about the bridge, seems to find sets of sensors that are in close proximity of each other. This makes sense, since signals travelling through the concrete are likely to diminish over the distance travelled.

Currently, we have experimented with various types of equations and search strategies, which all lead to similar sets of equations. Future research is needed to show if these assumptions hold when more data is considered, and in addition, a proper quality measure should be developed in order to evaluate such a minimalistic system.

3.3 Grid Solutions

In order to manage the magnitude of data that is associated with InfraWatch, one cannot rely on most existing data mining approaches. Most methods require either a reasonable part of the data to be in-memory or to be repeatedly scanned on disc. This makes the application of even trivial methods very hard on terabytes of data, as is the case in InfraWatch. To this end, one approach

is to apply *grid* techniques that can distribute one huge analysis task over thousands of computing nodes in order to make the tasks more feasible.

The application of grid techniques to data mining is not new. For example, a recent initiative, Weka4WS, has ported the well known Weka data mining platform to grid architectures [Talia et. al. 2005]. However, the nature of most of our analysis problems would require algorithms that are not readily available in Weka.

Although grids are well-suited to crunch large collections of data, they do this in a certain fashion. Many grid-enabled algorithms break down the dataset in small fragments, which are forwarded to the individual nodes for detailed analysis. However, in many data mining problems, many possible hypotheses have to be evaluated on *all* of the data. How to develop grid-enabled algorithms that can cope with our type of streams is a challenge to be addressed.

4 CONCLUSION

In this paper we have introduced the InfraWatch project, which is an infrastructure monitoring project centred on a highway bridge in the Netherlands: the Hollandse Brug. We have outlined the background of this project and infrastructure setting, and discussed data management issues on the bridge and on infrastructural assets in general. Our current focus is on making the analysis of the gathered data feasible, as the involved data volumes are huge. After discussing how these data volumes can be successfully managed, we discuss how and what type of analysis we see fit for our current scope.

5 ACKNOWLEDGEMENT

The authors are very grateful to the alliance ‘Hollandse Brug’ constituted of the Directorate-General for Public Works and Water Management (Rijkswaterstaat), Strukton Civiel and Reef Infra for their financial support and their contribution to the project which was essential for the establishment of the monitoring project and the experimental program.

6 REFERENCES

Aggarwal, C. C. 2007. *Data Streams: Models and Algorithms*, *1th Edition*, Springer

Boresi, A., Schmidt, R., and Sidebottom, O. 1993. *Advanced Mechanics of Materials, 5th Edition*. New York: John Wiley & Sons, Inc.

Rabinovich, O. and Frostig, Y. 2000. Closed-form higher order analysis of RC beams strengthened with FRP strip. Journal of Composite Construction, ASCE, 4: 65-74.

Teng, J., Zhang, J., and Smith, S. 2002. Interfacial stresses in reinforced concrete beams bonded with a soffit plate: a finite element study. *Construction and Building Materials*, 16: 1-14.

Koopman, A., Knobbe, A., and Meeng, M. Pattern Selection Problems in Multivariate Time-Series using Equation Discovery. in KDD'10 workshop - Useful Patterns.

Mueen, A., Keogh, E., Zhu, Q., Cash, S., Westover, B. 2009. Exact discovery of time series motifs. In: SIAM International Conference on Data Mining (SDM09). American Statistical Association (ASA).

Mueen, A., Keogh, E. Online Discovery and Maintenance of Time Series Motif, In the Proceedings of ACM SIGKDD 2010. pp. 1089-1098

Patel, P., Keogh, E., Lin, J., Lonardi, S. (2002) Mining motifs in massive time series databases. In: IEEE international conference on data mining

Talia, D., Trunfio, P., Verta, O. Weka4WS: a WSRF-enabled Weka Toolkit for Distributed Data Mining on Grids. In Proceedings of the PKDD 2005.

Todorovski, L. and Dzeroski, S. Declarative Bias in Equation Discovery, Proceedings of the Fourteenth International Conference on Machine Learning, 1997

Žliobaite, I. (2010). *Adaptive Training Set Formation*. Vilnius University, Lithuania.