Available online at www.sciencedirect.com

ScienceDirect

IFAC-PapersOnLine 49-20 (2016) 037–042

Mining Sensor Data in Larger Physical Systems

Paul O'Leary, Matthew Harker, Roland Ritt, Michael Habacher, Katharina Landl, Michael Brandner

* *Institute for Automation, University of Leoben, Leoben, Austria. (e-mail: roland.ritt@unileoben.ac.at).*

Abstract: This paper presents a framework for the collection, management and mining of sensor data in large cyber-physical systems. Particular emphasis has been placed on mathematical methods, data structures and implementations which enable the real-time solution of inverse problems associated with the system in question. That is, given a system model, to obtain an estimate for the phenomenological cause of the sensor observation. This enables the use of causality, rather than mere correlation, when computing measures of significance during machine learning and knowledge discovery in very large data sets. The model is an abstract representation of a real physical system establishing the relationships between cause and effects. The pertinent behaviour of the model is captured in the form of equations, e.g., differential equations. The inverse solution of these model-equations, within certain constraints, permit us to establish the semantic reference between the sensor observation and its cause. Without this semantic reference there can be no physically based knowledge discovery.
Embrechts pyramid of knowledge is addressed and shown that it will not suffice for future developments. The issue of information content is addressed more formally than in most data mining literature. Additionally the Epistemology for the emergent-perceptive portion of speech is presented and a prototype implementation with experimental results in data mining are presented. A lexical symbolic analysis of sensor data is implemented.

Keywords: Data mining, entropy, linear differential operators, lexical analysis.

1. INTRODUCTION

This paper addresses issues involved in mining sensor data in cyber physical systems (CPS), see e.g. O'Leary et al. (2015b). The advent of cyber physical systems has brought about a significant change in the architecture of sensor and measurement systems. The change is in terms of the numbers of sensors involved, the complexity of the system being addressed, the spatial distribution of the sensors within the systems and the global nature of the system itself. This is facilitated by the use of network components.

There are many different definitions for what constitutes a cyber physical system (CPS) Baheti and Gill (2011); Geisberger and Broy (2012); IOSB (2013); Lee (2008); NIST (2012); Park et al. (2012); Spath et al. (2013a,b); Tabuada (2006). The most succinct and pertinent to this paper is the definition given be the IEEE Baheti and Gill (2011) and ACM[1]:

A CPS is a system with a coupling of the cyber aspects of computing and communications with the physical aspects of dynamics and engineering that ***must abide by the laws of physics****. This includes sensor networks, real-time and hybrid systems.*

The solutions computed from the sensor data **must** obey the equations modelling the physics of the system being observed - this is fundamentally an inverse problem and requires the modeling of the system dynamics. Unfortunately, the issue of inverse problems is not addressed in literature on *mining sensor data*, see for example Esling and Agon (2012); Fuchs et al. (2010); Keogh and Kasetty (2003); Last et al. (2004). Actually, none of the standard books available, e.g. Aggarwal (2013) on mining sensor data, take the special nature of the sensor data into account. Present data mining techniques rely on correlation (in some manner) as being a reliable measure for significance. However, the inverse solution of the model-equations is required to establish the semantic reference between the sensor observation and its cause. Without this semantic reference to causality there can be no physically based knowledge discovery. The main contributions of this paper are:

(1) A structured approach based on *data, information, hypothesis, evidence, truth and fact* is proposed and demonstrated.
(2) A mechanism for the solution of ordinary differential equations (ODE) is introduced. The numerical methods for the linear differential operators (LDO) have been developed in such a manner that they can be integrated into standard data mining environments such as Hadoop Shvachko et al. (2010); White (2009). This permits the implementation of embedded simulation (forward problems), the solution of inverse problems and regularizing computations.

[1] ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS) (iccps.acm.org)

2405-8963 © 2016, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2016.10.093

(3) The extension of the symbolic aggregate approximation (SAX) Keogh and Kasetty (2003) to non-linear intervals while maintaining the lower bound property. The approximation with series of symbols permits the use of regular expressions to perform symbolic searches in the real sensor data.

2. A STRUCTURED APPROACH TO DATA ANALYTICS/MINING

Embrechts et al. (2005) proposed the pyramid of data mining as shown in Figure 1, which was an extension of Ackoff's work Ackoff (1989). Embrechts offers no definitions for the terms *information, knowledge, understanding and wisdom* in his work, while Ackoff offers intuitive but rather nebulous inaccurate definitions. The pyramid and the terms used have positive connotations [2] ; however, they do not provide a scientific basis for mining sensor data.

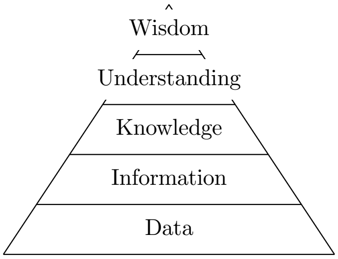

Fig. 1. The data mining wisdom pyramid as proposed by Mark Embrechts.

Shannon provided a formal mathematical framework for computing *information* content in a data stream. It is important to note that this measure of information is neither a measure for *significance* nor of *meaningfulness*. It is however a powerful tool in identifying temporal locations where the information content in a data stream changes.

The addition of *meta-data* is required to give a data stream *meaning*. For example, the stream of data from a digital thermometer is meaningless, i.e. its just a stream of numbers, without the meta-data for the measurement device. Correct meta-data should ensure that a *datum* is a statement about a *fact* and can so be considered as a low level of *knowledge*. In the context of this paper we shall define *understanding* as the relationship between facts, *constructs* (models) and *temporal context* which is indicative of a specific operation (*behaviour*) [3] . Understanding permits the prediction of the behaviour of a system under similar circumstances and *wisdom* can be considered as advantageous behaviour in a certain circumstance given the specific state of the available resources.

To implement data analytics and mining, it is necessary to develop tools which support the *exploration* of data which supports the formulation of *hypotheses*. Then we need a means of representing the hypotheses to enable the extraction of *evidence* from the data with the aim of supporting or refuting the hypothesis. This is the process of *knowledge discovery*. It requires a representation for knowledge which enables its later utilization — this mechanism must also include the possibility to encode a-priory knowledge, e.g., coming from the design process. *Statistical evidence* plays a very significant role in mining sensor data, since the complexity of the systems my preclude the complete *observability* of the process, i.e., there is no possibility to have a conclusive *induction* of the creating process from the observation represented by the sensor data in a finite time.

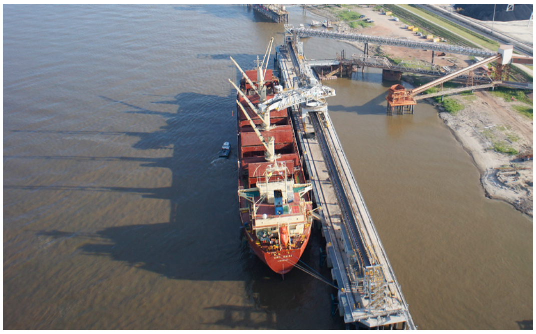

Fig. 2. Photograph of the port and ship loader used as an example to characterize the data flows involved in continuous monitoring.

3. DATA COLLECTION AND MANAGEMENT

Two modes of data collection are supported: *full-table* mode whereby all channels are provided at each time stamp. This mode is suitable for systems where there is a high degree of activity and no long periods during which the system is dormant; *on-change* in this mode values for the channels are only transmitted when they change, each message consists of a *time-stamp*, a *tag* and a *value*. In on-change mode it is necessary to introduce a calibration mechanism which initiates the transmission of all *time-stamp: tag: value triplets* at predefined times, otherwise there is the danger of undefined data states. At the server the sensor data is always stored in full-table format.

We have characterized the data flows from a ship loader currently being monitored, see Fig. 2. Currently, there are $n = 150$ sensors being monitored with the sampling period $t_s = 1s$, in addition there are four vibration channels acquired with $f_s = 2.5kHz$. The sensor data is transmitted as on-change and the vibration data as full-table, but segmented into observation periods of $t_p = 25s$. The summary of the data flows for this plant is given in Table 1. The volume of data transmitted varies from day to day due to the use of on-change format. The maximum file size observed over a time period of $m = 147$ days was $s_d = 62Mb$ and the average over this period was $35.2Mb$. we have chosen the HFD5 file format to store the sensor data, since in this manner the data can be efficiently stored as binary in double precision. In this work we are using MATLAB® for exploratory work and Python for established methods. The consequence of this investigation is, that approximately $s_y \approx 2.6\,Gb$ of storage is required, with level nine compression, to archive the complete sensor

[2] *Wisdom* just as the word *creativity* have positive connotations but resist any formal definition, see von Hentig (1998) for a discussion of this issue. Withour formal definition they do not form the basis for objective data analytics.

[3] We are using the word *behaviour* with some concern at this point, since it reflects the subjective behaviour of the operator. Large plant and machinery can generate data sequences which have a very strong operator dependency.