Optimizing Configuration Data using Prescriptive Analytics*

Alexander Wurl

Siemens AG Österreich, Corporate Technology, Vienna, Austria
alexander.wurl@siemens.com

Abstract. Suboptimal or erroneous configuration data in rail automation systems may cause serious safety and cost issues. This research work aims to address continuous optimization of such data by (i) connecting configuration and operation data by integration of heterogeneous data sources, and by (ii) application of prescriptive analytics methods to propose decision options on how to correct and optimize configuration data.

Keywords: Configuration Process, Data Integration, Asset Management, Prescriptive Analytics

This document contains the PhD research plan and is organized as follows. In Section 1 we define what will be accomplished by introducing our research questions. Section 2 presents the background knowledge. In Section 3 we explain the significance of the research contribution. Then in Section 4 we describe the methods adopted in the research. Finally, in Section 5 we present intermediate results and future work.

1 Research Questions

In industrial and infrastructural systems, like rail automation, product configuration is the activity of engineering and customizing a product to meet the needs of a particular customer. The product in question may consist of mechanical parts, services, and software - each with various parameters and properties that reflect the variability. The result of a configuration process are the *configuration data*, i.e., a digital model of the system specifying all details for installation, operation and maintenance of the facility. In the case of a rail automation system, configuration data contain, e.g., the bill of material, detailed configuration plans of all hardware parts, the station and track topology, the screen layout of the operator terminals, parametrization of the control software, etc.

While configuration data is specified at engineering time, *operation data* is continuously generated by trains and the interlocking and control systems at operation time. The amount of operation data generated each day is huge, because

* This work is funded by the Austrian Research Promotion Agency (FFG) under grant 852658 (CODA).

it contains all positions and speeds of all trains, as well as logs of all telegrams exchanged by the different subsystems.

The combination of these two data models - configuration data and operation data - allows for a feedback loop which has the potential to detect hardware defects or errors in the configuration data. Anomalies and unexpected behavior in operation data can be detected by statistical methods like principal components analysis or discriminant analysis. Error causes can only be detected by locating the corresponding configuration objects in the configuration data. This promising setting enables to build a prescriptive analytics framework [1] including various statistical analysis methods to explain the observed behavior and to propose decision options on how to modify the configuration data.

Another interesting application enabled by the availability of configuration and operation models is *predictive asset management*, where prediction models for the obsolescence of the various hardware parts are computed from different heterogeneous data sources. Configuration and operation data are complemented here by sales and order forecast models and contextual data, like weather data.

Based on these challenges, the following research questions will be tackled in the proposed dissertation thesis:

- Which data integration processes are suitable for preparing configuration and operation data for prescriptive analytics applications?
- Which sequences of statistical methods are appropriate for predictive asset management and anomaly detection in configuration and operation data?
- How can new or modified configuration rules/constraints be derived by prescriptive analytics methods in order to optimize configuration data?

2 Background

Beyond the application of mere statistical analysis methods, data analytics methods require a federated architecture of descriptive, predictive, and prescriptive analytics in combination with data models and a data warehouse [2]. To achieve reasonable results in analytics, ensuring data quality in the process of data integration is an inevitable prerequisite [3]. Bridging the gap of heterogeneous data sets, we aim at defining a data schema for both operation and configuration data which can be realized by providing a data model that accepts all properties of heterogeneous data sets. Since similar data implies various representations, the interchange of data between operations and configuration models are important tasks, i.e. the data scheme strongly relates to the resulting data quality. Despite of a lot of important efforts, model interoperability is still a challenging task, leading most often to hand-crafted bilateral integration solutions [4], suffering from high maintenance overheads, technology dependence, and scalability problems. Therefore, previous results regarding data integration in schema-based approach [5] shall be extended within the course of the proposed dissertation.

Data analytics in rail automation gains more and more interest [6]. Applying data analytics on configuration data, we intend to use techniques from data mining, machine learning, and anomaly detection. These techniques enable to examine large data sets to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. In more detail, considering data from various integrated data sets methods from multivariate analysis [10] serve as basis for our data analysis, i.e., statistical models capture relationships among many factors to allow the assessment of information which has significant impacts in trend predictions.

Following the statistical results of descriptive and predictive analytics, we believe in applying prescriptive analytics methods to propose optimal decision options for optimizations in product configuration. Basically, configuration can be defined as a "special case of design activity, where the artefact being configured is assembled from instances of a fixed set of well-defined component types which can be composed conforming to a set of constraints" [11]. As prescriptive analytics has evolved as a new research field in industry which focuses on describing the courses of actions and shows the influence of each action [12–14], there is a great potential to apply this concept for constraints that represent technical restrictions, restrictions related to economic aspects, and conditions related to production processes. Analytics methods are able to capture all related configuration data which contribute to a wider analysis of restrictions and can therefore show potential optimization options.

3 Significance

Data analytics is a highly active research field which has been driven mainly in business applications in the last decade. Recently, more and more industrial applications are adapting these methods, e.g., for sensor analysis for predictive maintenance. We explore the links between data analytics technologies and product configuration. The integration of configuration data in the process of analyzing operation data enables an easy localization and therefore a better understanding of the source of an anomaly.

In this work we contribute to solutions for two crucial problems in the domain of rail automation at the company Siemens AG Österreich: (i) predictive asset management and (ii) prescriptive analytics for correcting/optimizing railway engineering data. In the first scenario, forecasts of the form "How many assets of type A will be needed within the next N years?" are computed. Reliable forecasts are very important for guaranteeing the availability of all necessary modules in the future and allow for a solid version and lifetime management of module variants. The second scenario - the continuous correction and optimization of configuration data - is an important building block of guaranteeing safe and high-performance train operation.

The framework developed within the proposed dissertation goes beyond the rail automation use case. The methods developed are of general nature and may be adapted and applied to other industrial fields such as industry automation, power plants, or energy management. This is novel and highly promising, especially in the context of Smart Production and Industry 4.0.

4 Research design and methods

According to the research questions, we design a framework of methods with the following contributions.

1. **Heterogeneous Data Model Integration.** We need to integrate various data sources of different formats, like Excel and XML. As different business units use different tools and formats to maintain data, integration of data is challenging and prone to errors. Existing data quality methods fall short of a generalized approach that covers such a variety of data types in the domain of rail automation. Our contribution extends the notion of *signifiers* for a robust and at the same time typo-tolerant identification of objects of different sources [15].
2. **Multivariate Data Analysis.** Multivariate statistics are eminently suitable for anomaly detection and prediction trends [9]. Therefore, we develop techniques to extract statistical information and anomalies from the operation data. E.g., are there any hardware models or interfaces which frequently reboot? Have trains of different vendors different driving behaviour? These analyses will also integrate configuration and contextual data.
3. **Feedback from Operation to Configuration Data.** Configuration models represent all the different HW and SW element types along with their structure and constraints to build a system (e.g. a rail automation system). By following the statistical results of prediction, we believe that new rules or constraints can be learned by using, e.g., classification and regression tree methods to improve the configuration models. For example, certain types of modules may cause overheating if located next to each other in the hardware rack. External, contextual data, usually available as linked open data, may also be integrated to derive additional rules (e.g. heat sensibility of a module derived from module shutdowns in combination with meteorological data).

5 Research stage

The project related to this research has been started in April 2016. The work follows the contributions described in Section 4 assuming that they build on each other.

The first stage, Heterogeneous Data Model Integration, is finished. The result of the this contribution is an approach “Using Signifiers for Data Integration in Rail Automation” which was presented at the *6th International Conference on Data Science, Technology and Applications* [15]. This approach enables a semi-automatic process for data import, where the user resolves ambiguous data classifications. We introduced a technique using a *signifier*, which is a natural extension of composite primary keys to find the correct data warehouse classification of source values in a proprietary, often semi-structured format. This approach is already in use and results show a significant improvement of data quality.

The different data analytics tasks for predictive asset management and anomaly detection in operation data are defined and documented in a user requirements specification. Next, we will study the applicability of different multivariate methods to our analytics tasks. The selection and application of statistical methods is a highly sensitive task since the results serve as basis for further prescriptive analytics methods to optimize rules and constraints in product configuration.

References

1. Maglio, P.J.H.P.P., Selinger, P.G., Tan, W.C.: Data is dead without what-if models. Proceedings of the VLDB Endowment **4**(12) (2011)
2. Soltanpoor, R., Sellis, T.: Prescriptive analytics for big data. In: Australasian Database Conference, Springer (2016) 245–256
3. Bleiholder, J., Naumann, F.: Data fusion. ACM Computing Surveys (CSUR) **41**(1) (2009) 1
4. Schürr, A., Dörr, H.: Introduction to the special sosym section on model-based tool integration. Software and Systems Modeling **4**(2) (2005) 109–111
5. Papadakis, G., Alexiou, G., Papastefanatos, G., Koutrika, G.: Schema-agnostic vs schema-based configurations for blocking methods on homogeneous data. Proceedings of the VLDB Endowment **9**(4) (2015) 312–323
6. Rapolu, B.: Focus: How big data is making tracks in the rail industry. Building the Digital Transport Network of the Future (2015)
7. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier (2011)
8. Bishop, C.M.: Pattern recognition. Machine Learning **128** (2006) 1–58
9. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM computing surveys (CSUR) **41**(3) (2009) 15
10. Esbensen, K.H., Guyot, D., Westad, F., Houmoller, L.P.: Multivariate data analysis: in practice: an introduction to multivariate data analysis and experimental design. Multivariate Data Analysis (2002)
11. Sabin, D., Weigel, R.: Product configuration frameworks-a survey. IEEE Intelligent Systems and their applications **13**(4) (1998) 42–49
12. Souza, G.C.: Supply chain analytics. Business Horizons **57**(5) (2014) 595–605
13. Porter, M.E., Heppelmann, J.E.: How smart, connected products are transforming companies. Harvard Business Review **93**(10) (2015) 96–114
14. Siksnys, L.: Towards prescriptive analytics in cyber-physical systems. Dissertation (2014)
15. Wurl, A., Falkner, A., Haselböck, A., Mazak, A.: Using signifiers for data integration in rail automation. In: Proceedings of the 6th International Conference on Data Science, Technology and Applications - Volume 1: DATA,, INSTICC, SciTePress (2017) 172–179