In: Proceedings of the International UIS Workshop.
ITeG, University of Kassel, Germany. 2015

# The FEE Project: Introduction and First Insights

Andreas Schmidt, Martin Atzmueller and Gerd Stumme

Knowledge and Data Engineering Group,
Research Center for Information System Design,
University of Kassel, Germany
{schmidt,atzmueller,stumme}@cs.uni-kassel.de

## Introduction

In many industrial areas, production facilities have reached a high level of automation nowadays. On the one hand, this is crucial for an economic operation of the production facilities, especially in countries with high salaries like Germany. On the other hand, process automation has some drawbacks for the facility operator. When the production facility is running in automatic mode most of the time, the facility operator can not gain experience about the dynamic characteristics of the production facility which can lead to a decrease in experience with the production process. Know-how about the production process is crucial, especially in case the production facility reaches an unexpected operation mode such as a critical situation. When the production facility is about to reach a critical state, the amount of information (so called shower of alarms) can be overwhelming for the facility operator, eventually leading to loss of control, production outage and defects in the production facility. This is not only expensive for the manufacturer but can also be a threat for humans and the environment.

Therefore, it is important to support the facility operator in a critical situation with an assistant-system using real-time analytics and ad-hoc decision support. The objective of the BMBF funded research project "Frühzeitige Erkennung und Entscheidungsunterstützung für kritische Situationen im Produktionsumfeld"[1] (short FEE) is to detect critical situations in production environments as early as possible and to support the facility operator with a warning or even a recommendation how to handle this particular situation. This enables the operator to act proactively, i. e., before the alarm happens, instead of just reacting to alarms.

The contribution of this workshop paper is to share our experiences and first insights about the preprocessing of an industrial, real world dataset that was provided by one of our application partners and to describe first experiments performed on this dataset.

[1] http://www.fee-projekt.de

## Overview

The consortium of the FEE project consists of several partners including also application partners from the chemical industry. These partners provide use cases for the project and background knowledge about the production process which is important for designing analytical methods.

The available data was collected in a petrochemical plant over many years and includes a variety of data from different sources such as sensor data, alarm logs, engineering- and asset data, data from the process-informations-management-system as well as unstructured data extracted from operation journals and operation instructions. The heterogenous data will be consolidated and integrated by a central big data analytics platform (see Figure 1). Analytical methods will be built on top of this platform and make use out of the heterogeneous mass data with regard to big data capabilities. This can be an early warning (e. g., likelihood alarm) or an assistant function that gives a recommendation how to handle a specific situation. Furthermore feedback from the facility operators will be used to continuously improve the assistance function.

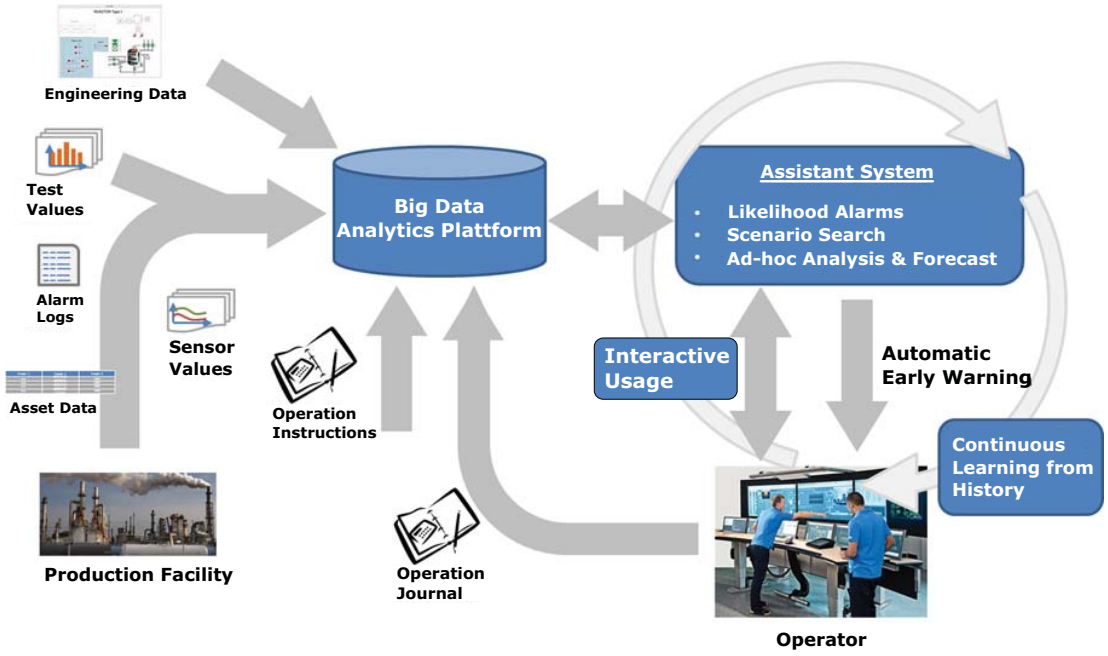


**Figure 1: The big data analytics plattform consolidates and integrates heterogenous mass data collected over many years. The assistant system is build on top of the plattform and utilizes analytical methods which automatically generate an early warning or even give a recommendation how to handle a specific situation.**

## Experiences and First Insights

The dataset consists of various different document types. Unstructured / textual data is included as part of the operation instructions and operation journals. Knowledge about the process dependencies is provided as a part of cause-effect-tables. Information about the production facility is included in form of flow process charts. Furthermore there is information about alarm logs and sensor values coming directly from the processing line.

One key aspect prior to an analysis is the anonymisation of the data. To protect personal data of individuals from further analytics, person names should be made irrecognizable. One way to achieve this is to simply remove those names from the document. This follows the principle "privacy by design" which means that anonymisation should be performed as early as possible, so that person names can not be the subject of further analytics any more. For automated anonymisation it is also a requirement to convert the files from binary to text formats. Note, that such a conversion can cause a loss of information. In case of graphical documents, for example, it is obvious that the visual information cannot be captured by a text format. Such documents have to be manually anonymised by hand (e. g., blackening person names). For text processing documents (e. g., word format) most of the information can be preserved by choosing an html format over a plain text format. This way, the document structure (e. g., headlines, bold words) is still available for processing such as generating warnings for abnormal situations.

An assistant system that generates early warning for abnormal situations actually has to solve two types of tasks. At first, the system has to identify events based on the long history of data. A burst in the frequency of the alarm logs, for example, could be an indicator for an unexpected situation which corresponds to an event. Secondly, after identifying events, the system needs to extract features that help to predict this type of event as early as possible. Coming back to the frequency of alarm logs example, small fluctuations in the frequency distribution could be an indicator for a specific event characterized by a burst in the alarm log frequency. With the data, we performed first experiments; results of which will be presented at the workshop.

## Outlook

Future work will include the analysis of extraordinary characteristics in the industrial real world dataset. One algorithm that can be applied for exploratory analysis is Exceptional Model Mining [1, 10], as a variant of subgroup discovery [9, 13] focusing on complex target properties. For that, there are fast implementations available, e. g., as implemented in the VIKAMINE[2], cf. [3, 4], and the *rsubgroup* package[3] for the *R* platform for statistical computing.[4]

Furthermore we plan to apply techniques from information extraction [8] to the unstructured / textual information for event detection [11], e. g., using rule-based techniques [2]. By applying NLP techniques, we will analyse the potential of extracted information for indicating upcoming events and for obtaining analytical profiles, e. g., [6]. Then, case-based and network-based techniques, e. g., [5, 12] can provide further options for analysis and explanation, cf. [7].

[2] http://www.vikamine.org
[3] http://www.rsubgroup.org
[4] http://www.r-project.org

## Acknowledgement

This work was funded by the BMBF project FEE under grant number 01IS14006E.

## References

[1] M. Atzmueller. Subgroup Discovery – Advanced Review. *WIREs: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015.

[2] M. Atzmueller, P. Kluegl, and F. Puppe. Rule-Based Information Extraction for Structured Data Acquisition using TextMarker. In *Proc. LWA 2008*. University of Wuerzburg, 2008.

[3] M. Atzmueller and F. Lemmerich. VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In *Proc. ECML/PKDD*, 2012.

[4] M. Atzmueller and F. Puppe. SD-Map – A Fast Algorithm for Exhaustive Subgroup Discovery. In *Proc. PKDD 2006*, number 4213 in LNAI, pages 6–17, 2006.

[5] M. Atzmueller and F. Puppe. A Case-Based Approach for Characterization and Analysis of Subgroup Patterns. *Journal of Applied Intelligence*, 28(3):210–221, 2008.

[6] M. Atzmueller, F. Puppe, and H.-P. Buscher. Profiling Examiners using Intelligent Subgroup Mining. In *Proc. IDAMAP-2005*, pages 46–51, Aberdeen, Scotland, 2005.

[7] M. Atzmueller and T. Roth-Berghofer. The Mining and Analysis Continuum of Explaining Uncovered. In *Proc. 30th SGAI International Conference on Artificial Intelligence (AI-2010)*, 2010.

[8] R. Grishman. Information extraction: Techniques and challenges. In *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. Springer, 1997.

[9] W. Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*, pages 249–271. AAAI, 1996.

[10] D. Leman, A. Feelders, and A. J. Knobbe. Exceptional model mining. In W. Daelemans, B. Goethals, and K. Morik, editors, *ECML/PKDD (2)*, volume 5212 of *Lecture Notes in Computer Science*, pages 1–16. Springer, 2008.

[11] G. Melton and G. Hripcsak. Automated detection of adverse events using natural language processing of discharge summaries. *J Am Med Inform Assoc*, 12(4):448–57, 2005.

[12] F. Mitzlaff, M. Atzmueller, D. Benz, A. Hotho, and G. Stumme. Community Assessment using Evidence Networks. In *Analysis of Social Media and Ubiquitous Data*, volume 6904 of *LNAI*, 2011.

[13] S. Wrobel. An Algorithm for Multi-Relational Discovery of Subgroups. In *Proc. PKDD*, volume 1263 of *LNCS*, pages 78–87, 1997.