Human Aspects of Advanced Manufacturing, Production Management and Process Control, Vol. 184, 2025, 21–26
https://doi.org/10.54941/ahfe1006446

Complexity Reduction Using AI-Based Correlation Analysis Using the Example of Operator Actions

Wolfgang Krause[1] and Daniel Schilberg[2]

[1]AWG Abfallwirtschaftsgesellschaft mbH, Wuppertal, NRW, Germany
[2]University of Applied Sciences Bochum, Bochum, NRW, Germany

ABSTRACT

This Paper examines the use of artificial intelligence (AI) to analyse and optimise operating messages in a waste-to-energy plant. The aim of this work is to identify potential for improvement in automation technology. For this purpose, the 13,9 TB of data collected over the years from the message archive of the Wuppertal waste-to-plant are analysed, in particular the operating messages. The Siemens® Simatic PCS7 Process control system, which has been in use since 2016, is used as the data source. An MSSQL server is set up on a Linux system to import these files and make them accessible. With MATLAB as a client and using the Statistics and Machine Learning Toolbox, AI algorithms are used to analyse correlations between the operator messages. The aim is to recognise patterns and connections that indicate inefficient processes or unnecessary messages. The results of this analysis will be used to optimise the automation technology. This will reduce the workload for operators at the control centre and at the same time increase the efficiency and safety of power plant operation.

Keywords: AI, Production, Process management

INTRODUCTION

Waste-to-energy plants play a crucial role in modern waste management, not only by generating energy for electricity and district heating networks, but also by acting as a pollutant sink and making a decisive contribution to the environment by hygienising the waste. Such plants require precise control, regulation and monitoring to ensure optimum operating conditions and minimise environmental risks. Although advanced process control systems such as Siemens® Simatic Process Control System (PCS)7 have improved the monitoring, control and regulation of these plants, operators are faced with the challenge of organising the flood of messages efficiently and effectively for the operators. This challenge is exacerbated by two main factors. Demographic change: The shortage of skilled labour due to the retirement of baby boomers will be exacerbated by low numbers of successors. Between 2024 and 2040, there will be a shortage of 3.5 million people on the labour market (Destatis, 2024). This will make it more difficult to fill key positions and increase dependence on existing staff. Knowledge management: When

the baby boomer generation retires, valuable expertise is lost. Building up expertise requires time, which is becoming scarce due to demographic trends. Without effective strategies, there is a risk of losing company knowledge in the long term (Thim, 2012).

METHODOLOGY: STRATEGIES FOR DATA PROVISION AND ANALYSIS

This part describes the methodological approaches of the work, including data provision, the database technologies used, the analysis tools employed and artificial intelligence (AI) algorithms. The focus is on the efficient structuring and analysis of the extensive reported data in order to optimise operational processes.

Database Technology and Data Structure

The reports are stored and organised using an Microsoft® Structured Query Language (MSSQL) server that is implemented within Simatic PCS7. The reports are stored in segmented archives. Each segment is composed of two file types. The files with the extension.mdf contain the main data, while the files with the extension.ldf serve as transaction logs to ensure data integrity. The segmented architecture allows for the automatic archiving of older data if the defined storage limits or time periods are exceeded. The central tables within the segmented tables include, for example, MsArcLong (main table for archived reports) and AlgCSDataDEU (contains configuration-specific message texts in German) (Murugesan, 2015).

Integration of the MSSQL Server

A separate MSSQL server was implemented to provide the message data. This is based on a Dell® PowerEdge R660 that has been specially configured for performance intensive applications. With two Intel® Xeon processors, 256 Gigabyte (GB) of RAM and an NVMe SSD (7.68 Terabyte (TB)), the system meets the requirements for extensive data processing. The archive segments were integrated into the MSSQL server by importing the .mdf and .ldf files. This architecture provides the basis for a seamless connection to the analysis system. In addition, the network infrastructure was optimised to enable fast communication between the MSSQL server and the analysis systems.

Analysis With MATLAB and AI Algorithms

Both numerical and textual analyses were carried out using the toolboxes available in Matrix Laboratory (MATLAB). The Statistics and Machine Learning Toolbox supported pattern recognition in numerical data, while the Text Analytics Toolbox enabled complex text analyses. A central tool for the analysis was the Bag-of-Words-Modell (BoW), which transforms text data into frequency vectors. This made it easier to analyse and compare the messages. The method was extended by the Bag-of-N-grams-Modell (BoN) to the effect that not individual words but sequences of words, so-called N-grams, were considered. This proved to be particularly useful for identifying recurring phrases or patterns in the messages. The mentioned models served as a basis for machine learning methods, which enabled an

analysis of typical language patterns and their variations (Math1, 2024 and Math2, 2023). Furthermore, Latent Dirichlet Allocation (LDA) was used to identify latent topics in the data. The generative model assigns messages to thematic categories based on the frequency of certain words. The application of LDA proved to be particularly effective for identifying patterns in large amounts of data and analysing their correlation with operational processes (Blei, 2003). For example, topics that occurred more frequently during certain shift times could be identified. This enabled conclusions to be drawn about time-dependent operating conditions. Another essential tool for the analysis was the creation of co-occurrence networks. The relationships between the terms in the reported data were visualised by representing the words as nodes and the common frequency of the words as edges. Analysing the data using these networks allowed for intuitive exploration and efficient identification of relationships. For example, the terms frequently associated with the term "button" were examined. This allowed conclusions to be drawn about recurring causes. The application of such networks proved particularly useful for revealing rare but significant relationships in large amounts of data. The combination of these tools made it possible to uncover both statistical patterns and semantic relationships in the reported data. This provided the basis for well-founded optimisation measures in the operation of the plant. The present scientific work was written from the perspective of an engineer and not from that of a computer scientist. Accordingly, the focus is on the implementation of a software that not only fulfils the tasks required for data processing, but also provides ready-to-use algorithms. The use of this software does not require in-depth knowledge of programming or the development of algorithms. These requirements are fully met by MATLAB.

DATA ANALYSIS: INVESTIGATION AND INTERPRETATION OF THE OPERATIONG DATA

The data analysis is carried out in several phases that build on each other, with each phase contributing to the acquisition of well-founded insights, see the Figure 1.

The process begins with selection, in which the data that is relevant for further analysis is identified and chosen (Cleve, 2016). This section is the

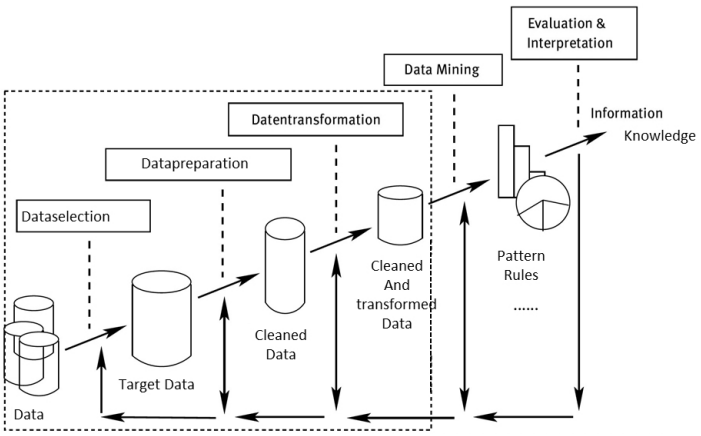

Figure 1: Process flow in data mining (Cleve, 2016).

actual Structured Query Language (SQL) query in MATLAB code, which is used to retrieve the data from the SQL server. In the data preprocessing phase, the raw data is cleaned up. For this purpose, missing values are supplemented by suitable replacement values and incorrect or contradictory data is corrected. This ensures a consistent and reliable basis for the subsequent analysis steps. In the case of reported data, comprehensive data preparation is not necessary because it comes from standardised and automated processes, which ensures a high level of data quality. Simatic-PCS7 is designed to generate complete data. Furthermore, reported data is immediately monitored by the operating personnel and anomalies are corrected. The data is then transformed, whereby the data is converted into adequate data formats. In many cases, the methods used in the data analysis require the use of specific data formats. Consequently, the data is transformed in this phase so that it meets the requirements of the subsequent analysis procedures (Cleve, 2016). In MATLAB, the automatic data type recognition in SQL queries leads to inaccurate results. To avoid this, the data types for each column are defined manually to ensure correct conversion. In some cases, columns were also merged after the query.

The core of the data analysis lies in data mining, where models such as decision trees are developed to uncover patterns and regularities in the data. This phase is crucial to identifying and modelling valuable relationships. For the text-based analysis, the BoW and BoN models were used (Math1, 2023; Math2, 2023). These models identified frequently used terms and phrases. A cleaned word cloud clearly showed the most frequently occurring terms. Furthermore, n-grams made it possible to analyse the relationships between words. The LDA model identified topics within the data, with 14 topics being generated. These topics were displayed as word clouds and bar charts, which helped to show relationships such as the high correlation. Another focus was the creation of co-occurrence networks that visualised the relationships between terms. For example, the network around the word "Button" showed how it is associated with terms such as "Acknowledged" and specific plant components. These networks provided valuable insights into operational dependencies and allowed conclusions to be drawn about possible process optimisations. The interpretation and evaluation of the results then follows. As part of the evaluation, the patterns found are examined to determine whether they are novel, useful and practically applicable. The analysis revealed clear correlations between events and terms that indicate weaknesses. One example was the venting valve, which could be optimised by automatic control based on level measurements. This shows that the identified patterns are not only statistically relevant, but also have practical applications for increasing the efficiency and safety of the plant.

The analysis carried out showed that a comprehensive investigation of all relevant operating messages requires an extremely large amount of time, since the computational effort for the Central Processing Unit (CPU) load is extremely high. The calculations carried out showed that the calculation of the data sets under optimal conditions and without further resource bottlenecks would take more than 29 days of pure CPU time. A complete recording and evaluation of all relevant correlations is associated with a

considerable expenditure of time, so that a complete analysis of all operating messages over a period of several years would be necessary. The results open up new perspectives for future applications. The implementation of real-time AI algorithms could enable a further increase in system performance and the establishment of automated decision-making processes. Transferring the methodology to other areas of waste management and environmental technology also promises promising results, for example with regard to optimising maintenance intervals. The creation of a central data pool from several plants could form the basis for big data analyses in the long term, supported by cloud technologies to further increase efficiency and scalability. In summary, the work provides valuable impetus for the further development of AI-supported solutions in the field of waste management and shows their relevance for practical application.

DISCUSSION: FINDINGS AND OPTIMISATION POTENTIAL

The main requirements were successfully implemented as part of the data analysis. The aim of the data analysis was to identify optimisation potential for user actions by systematically analysing the data with suitable algorithms. In particular, algorithms from the MATLAB environment were used as part of the data analysis. However, challenges also arose, in particular due to the high CPU load of the algorithms used, which included tokenisation, co-occurrence network analysis and LDA, among others. This resulted in extended processing times and delays because the hardware requirements were initially underestimated. Insufficient planning of resources had a negative impact on the analysis process, which highlights the importance of better aligning future projects with technical requirements. One potential solution would be to parallelise calculations or to use more powerful server structures to meet the high computing requirements. This could lead to an increase in efficiency and an acceleration of extensive data processing. In summary, it can be stated that the requirements have been largely met. However, the analysis has shown that the technical infrastructure and resource requirements must be given more consideration in future projects.

CONCLUSION

The work demonstrates that the use of AI -based text mining methods, such as co-occurrence networks, is an effective strategy for analysing and optimising process control engineering in Waste-to-energy plants. The identified patterns and correlations provide concrete approaches for reducing the workload of operating personnel and optimising operations management. The research question of whether AI can reduce complexity is confirmed. At the same time, it becomes apparent that the high time required for the analysis highlights further optimisation potential.

The defined goals and criteria were met, although a complete analysis of all operations was not possible due to the large amount of data. Consequently, a significant reduction in workload could not be demonstrated. The high data load and the long processing times require optimised programming, in

particular through debug functions and parallelisation, in order to increase the efficiency of the analysis.

ACKNOWLEDGMENT

The authors would like to acknowledge the AWG Abfallwirtschaftsgesellschaft mbH Wuppertal for the support.

REFERENCES

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, Eds. J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003. http://portal.acm.org/citation.cfm?id=944937

Cleve, J. and Lämmel, J. (2016). Data Mining, 2nd ed., ser. De Gruyter Studium. Berlin and Boston: De Gruyter Oldenbourg.

Destatis, (2024). Federal Statistical Office of Germany, https://www.destatis.de/DE/Im-Fokus/Fachkraefte/Erwerbstaetigkeit/_inhalt.html.

Math1 (2024). https://de.mathworks.com/help/textanalytics/ref/bagofngrams.html

Math2 (2023) I. The MathWorks, Text Analytics Toolbox™ User's Guide, I. The MathWorks, Ed. The MathWorks, Inc.

Murugesan, D. M. and Karthikeyan, K. (2015). Analyzing integral components of sql server databases," International Journal of Applied Engineering Research.

Thim, C. and Weber N. (2012). Demographischer Wandel - Herausforderungen für die Arbeits- und Betriebsorganisation der Zukunft. Gito Verlag GmbH, 09 2012, ch. Herausforderungen des Demographischen Wandels an den Transfer von Erfahrungswissen, pp. 361–382.