Knowledge-based Production Documentation Analysis: An Integrated Text Mining Architecture

Johannes Zenkert, Christian Weber, André Klahold and Madjid Fathi
University of Siegen
Institute of Knowledge Based Systems and Knowledge Management
Siegen, Germany
{johannes.zenkert, christian.weber, andre.klahold, madjid.fathi}@uni-siegen.de

Kai Hahn
University of Siegen
Micro System Design Group
Siegen, Germany
kai.hahn@uni-siegen.de

***Abstract*—Across the landscape of high technology companies, knowledge is a vital core resource at the heart of the organization. The technological complexity is steadily rising, tightening the global demands which are leading to an aggravating conflict on the product: it has to be produced faster, cheaper, more customer related and all without failures, while the change of complexity is inviting new unseen potentials for faults. On top, a new need for an agile production is arising. While todays organizational knowledge is captured and managed by an increasing number of processes and methodologies, solutions to continuously re-integrate the captured knowledge into the overall process of design and production are still scarce. This paper presents an integrated and agile process that proposes an integrative text mining architecture for design and process analytics.**

***Keywords*—Integrative Text Mining, Textual Process Analytics, Smart Production, Intelligent Document**

I. Introduction

To gain an advantage in high technology production markets, a conscious, detailed and pervasive tracking of product, product design and production related documents, as well as measurements and indicators is essential. Production failures can create critical production scenarios with the need for recovery or expensive redundancy, leading to rising costs, which on the longer run may shift the market position. Furthermore, improvements of state-of-the-art technology or even single processes are hard to reach without a steady access to information. So, beyond areas as e.g. failure prevention, the accessibility of information is a core enabler to fully utilize the intellectual capital of an organization and to uncover improvements which on short term lead to a better, less error prune and more informed production which then may lead to an improved and sustainable market position. To reach this stage, access to information is essential and, in case of existing information, information has to be used, reused, stored, maintained and rendered to be a "findable" resource.

To create a meaningful connection between the information within an organization and the different stages of the design and production process, smart and flexible solutions are needed to link the right information to the right processes and make them available on demand and on time as discussed in [1].

In recent years, text mining has established itself as a state-of-the-art method for the processing and analysis of unstructured textual data. Integrative Text Mining is a novel approach that combines multiple results from text mining methods into a holistic view of the data to cover different perspectives and semantic connections of individual analysis dimensions. In the area of design and process support for companies, document-oriented knowledge bases have become widespread as a support tool for knowledge intensive tasks and are used to store, manage, share, use and preserve organizational knowledge. However, solutions which interlink extracted information and processes are still scarce.

This paper discusses an architecture and its application to extract, describe and contextualize unstructured textual resources and represent the results in a way that they can be interlinked for different stages of the overall product life cycle. To do so, a set of distinct resource types is considered - documents, as the core information source; persons, as stakeholders and connectors; analytical feedback, as analytical cases for incidents which are stored in structured documents with a resource meta-structure. Further structured content, as database content, is highly relevant but will not be considered within this publication. Furthermore, this paper takes a data-driven, analytical perspective, focusing on the resources and how a semantic context can be extracted using text mining, and will not take a process-driven product life cycle management (PLM) perspective, which will be in focus for future works.

Section 2 discusses related work in the overall context of integrative text mining, knowledge representation and meta-data enriched documents. In Section 3, the knowledge-based architecture for the analysis of design and process documentation is introduced. Related components and methods are described in Section 4. Section 5 further outlines a four-step approach for the composition of intelligent documents based on textual process analytics. In the conclusion, the results of the paper are summarized, and future work is highlighted.

II. Related Work

To propose a novel architecture, different sources of information have to be considered for the information extraction process which is needed to tag annotated sources, together with related works in the areas of integrative text mining, knowledge representation and meta-data enriched documents.

978-1-5386-7392-8/18/$31.00

A. Integrative Text Mining & Knowledge Representation

Integrative text mining is a new field of application that combines the results of text mining methods into a common knowledge representation. Multidimensional Knowledge Representation (MKR) has been proposed in [2], [3]. For the presented architecture the methods named entity recognition (NER) [4], topic detection (TD), sentiment analysis (SA) and semantic triple extraction (STE) are used. With NER, the entities mentioned in the text - and thus relevant semantic relationships within the design and process documentation - are recognized, linked to respective documents and made available in the knowledge base. Named entities refer to designations for machines, manufacturing systems, product names, process components, names of experts and other knowledge assets referenced in the text. In TD, a distinction is made between document-oriented topic detection and multi-topic detection, which allows the text components to be assigned more precisely to subtopics. SA examines the polarity and opinions within a text and focuses on the positive or negative choice of words within a text, on document-, sentence-, entity- or aspect-level. During the extraction of semantic information, the use of language is further examined to generate facts from the written text based on the grammar of the text. All methods are either using frequency-based criteria in combination with dictionaries (e.g. term frequencies), co-occurrence-based analysis and evaluation (e.g. word association strength), or as classification models, trained by machine learning.

B. Meta-Data Enriched Documents

The high technization of production lines in high technology industries leads to increasingly complex environments. The increasing complexity correlates with a steadily rising number of potential sources of failure, which require a coordinated approach to failure analysis. While the overall process of failure analysis is well organized, utilizing the process of data mining [5], the documentation of analyzed failure cases is often captured in an unstructured manner - if documented at all. Several defined formats and also technologies to draft new structured formats exist, as XML, JSON, JAML or RDFS, which enable the needed human- and machine-specific readability. A step further are domain specific formats which are structured to support a specific scenario with full or limited extendibility and storing only results. An additional extension are formats which contextualize conducted experiments and model additional meta-data as e.g. persons, experiment descriptions and information which enable linking to other resources. With ChemKED, Weber and Niemeyer introduce a data-focused format for storing experiments and results in the chemical domain, including authoring information and sufficient formalization to simulate data of given experiments [6].

III. Knowledge-Based Architecture

The approach for analyzing the process and design documentation of the knowledge base is shown in Figure 1. As the overview indicates, multiple sources of knowledge and a specific set of methods are needed to capture a context-aware perspective which can interlink resources on a content, owner and process level. The architecture is conceptually designed to cover components which are universally needed to process resources, while specific cases may apply only a partial process with a specific parameterization. The architecture is describing a multistage process:

1) First, input information, concerning the overall product creation flow, processes and design is transferred to the knowledge base, including a) design information, as design manuals, client specifications and technological limitations; b) technical and relevant organizational processes, as quality management guidelines, capturing regulated well-defined and enforced production and quality ensuring procedures; and c) existing lessons learned from past and current production incidents, stored in semi-structured documents, utilizing standardized metadata.
2) This structured and unstructured information is then pre-processed, and natural language processing (NLP) operations are performed.
3) With the help of a process-relevant text corpora, text mining methods are then carried out. These methods are further trained and refined with the help of machine learning. Knowledge extraction within design and process documentation is carried out in particular by NER, TD, SA and STE. The process has already been described in related work [3]. The resulting Multidimensional Knowledge Representation (MKR) combines all relevant knowledge from underlying textual content.
4) With the help of intelligent search methods, including the knowledge inference or the MKR operations dimensional filtering, selecting and transforming, knowledge content is finally made available for visualization and knowledge discovery.
5) These facts are then enriched using visualization strategies and tailored text summarizations. This building blocks are then aligned into a smart document frame, the intelligent document [7], which on demand collects information into a readable format and offer the option to zoom deeper into specific information with inter-document links. Further lessons learned are integrated, incorporating analytical feedback from structured incident reporting.

IV. Components and Methods

The proposed architecture provides the essentially needed components for the contextualization and integration of the entire documented design and process knowledge into one MKR powered knowledge base. In this way, relevant documents are searched using metadata, extracted content and linked information, incorporating the organizational knowledge. The common knowledge base gives the company the advantage of working with consistent data and thus a uniform view of the organizational knowledge, interlinking with related processes and stakeholders in the company. The proposed

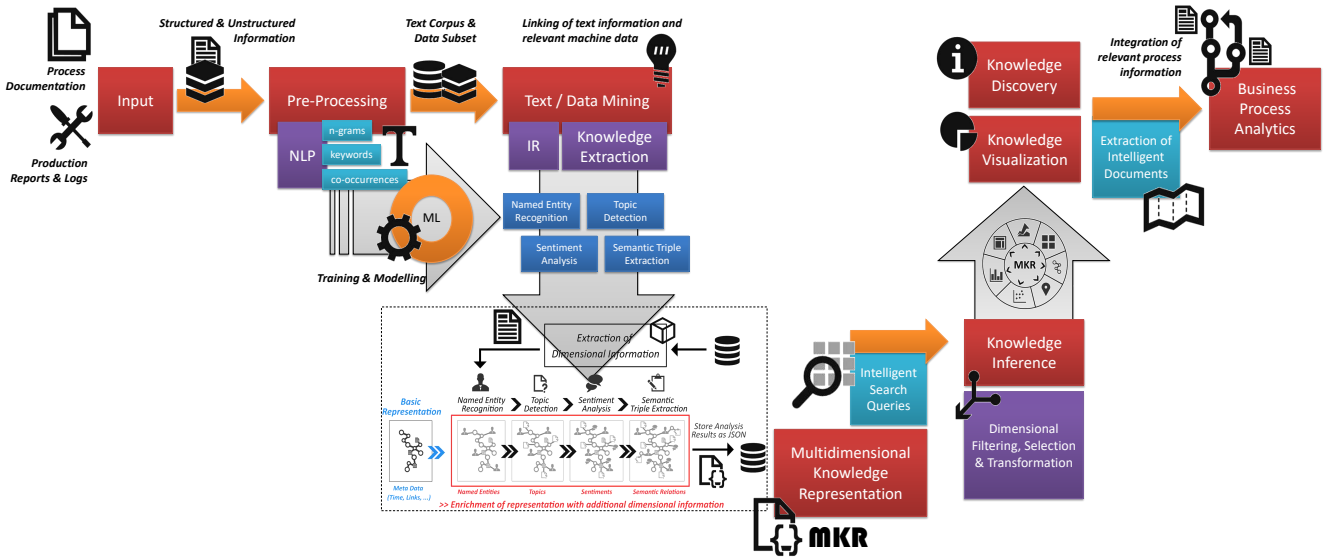

Fig. 1. Architectural Overview of Process Analytics through Integrative Text Mining utilizing Multidimensional Knowledge Representation

architecture implements implicitly five distinct methods across its components:

a) Mining for Information: Intelligent Search Methods: With the help of data and text mining methods, the documents are examined, and relevant content is extracted. Documents can contain either purely unstructured data, that is text, or they can already be implicitly linked to structured data, as e.g. in the case of tables. Search results provide an organized overview of requested information and enable consecutive complex methods

b) Multidimensional Knowledge Discovery within the Representation: In MKR, knowledge discovery is supported by the integration of dimensional information. Each dimension represents the pre-processed result of an individual method (NER, SA, TD, STE) which can be used as additional filter. Based on the availability of this additional information and without further calculation, adjustments are recommended for visualization or knowledge base searching.

c) Multidimensional Adaptive Text Summarization: Multidimensional text summarization is an adaptive process that combines and compacts textual information from individual documents or collections of documents, using selected dimension criteria and filters. Text summarization is generally utilized to shorten a text given a maximum number of text length. Here, through MKR, the text is adaptively adjusted according to information gained through individual results from text mining methods [8].

d) Visualization of the text-referenced knowledge items in process contexts: Through a graphical view on contextualized, linked knowledge items, facts are presented clearly and in context and can be examined by further calculations, e.g. based on centrality measures for visualizing the strength of association or the grouping of related topics [9].

e) Intelligent Documents: Intelligent documents are a smart medium for the summarization of design and process knowledge, enriched with relevant error reports, error analyses and additional excerpts from the knowledge base, to be considered for process, design or product changes. An intelligent document is utilizing the contextual background which has been extracted from textual information. The document integrates and organizes the content based on the relations, extracted and enriched by the MKR process.

V. A Multi-Level Process for Intelligent Document Composition based on Textual Process Analytics

Section 4 addresses components and methods which are provided by the proposed architecture. Each subsection is building upon the preceding method, finally creating the capacity for a range of different analytical outputs. The envisaged motivation is the creation of a compound knowledge base extract - the intelligent document - in order to answer or solve a specific question or problem. The intelligent document contains references to identified process documentations and relevant extracted sources of knowledge (including experts as explicit entities, to indicate implicit process knowledge). Furthermore, analytical results from previous incident analyses are included as so-called lessons-learned. For subsequent changes in product generations and processes, identified errors, resulting from incident analyses (e.g. log analysis) and causes of errors in production (machine and production reports), can be directly taken into account.

For the creation and enrichment of intelligent documents we propose a four-step approach illustrated in Fig. 2. Different knowledge assets and organizational knowledge are used in the architecture. For a clear distinction between identified and extracted assets, we consider the human-organization-technology (HOT) perspectives of knowledge management. In the human perspective we consider entities such as experts; the organizational part consists of process and design descriptions,

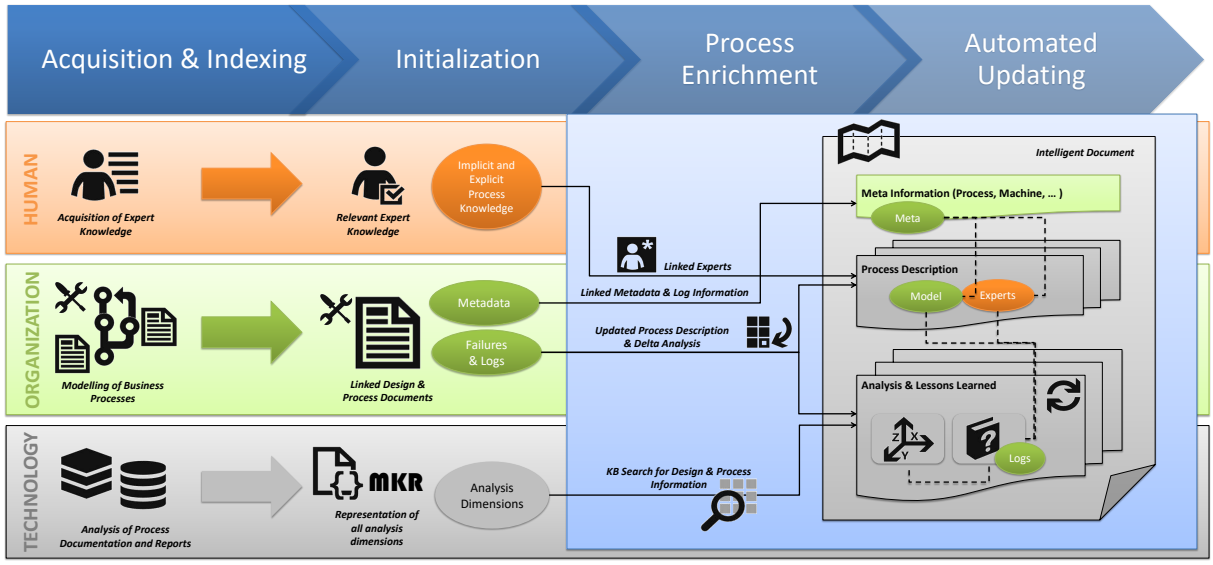

Fig. 2. A Multi-Level Process for Intelligent Document Composition based on Textual Process Analytics

incident reports and failure analysis; the linkage between related knowledge assets such as textual content of documents, logs or reports - identified and reflected by MKR - belongs to the technology perspective.

a) Phase I - Acquisition and Pre-Indexing: In the acquisition and indexation phase, all relevant knowledge assets - including experts and extracted content from the knowledge base - are considered. Furthermore, textual information from design and process documentation is identified.

b) Phase II - Initialization: In the initialization phase, relevant knowledge sources are gathered based on the acquisition and indexing results and metadata is assigned to documents. Based on each documents MKR, other entities and semantically-related content from other documents are referenced. Incident records are considered as complex documents which incorporate and link to multiple knowledge assets.

c) Phase III - Process Enrichment: During the process enrichment phase, relevant content is collected and finalized in the format of an intelligent document. The document is a composition of identified knowledge assets. Based on interaction, the document can be modified, further enriched by additional recommendation of assets.

d) Phase IV - Automated Update Cycle: The iterative, automatic enrichment of the intelligent documents is running continuously and automatically modifies or adds content to recommend new assets in the case of new or changed assets (e.g. similar metadata such as keywords, entity-linkages, related process contexts, new incident reports).

VI. Conclusion

This paper proposes an integrative text mining architecture to analyze process and design documentations, utilizing the MKR framework. Five components are mentioned for the contextualization and integration of the design and process documents into a knowledge base. The extraction of knowledge is further illustrated with the composition of an intelligent document which maps individual knowledge assets, entities and references to identified process documentations. The MKR text mining framework has been implemented in previous work. In future work, the remaining components of the proposed architecture will be added. The final objective is to integrate process-relevant knowledge into the form of an intelligent document. The architecture lays a foundation to the continuous use and re-use of knowledge in the design and production processes and will be a starting point for new knowledge-intense and agile applications, which are needed in the industry 4.0 context.

References

[1] R. Montino and C. Weber, "Industrialization of customized ai techniques: A long way to success!" in *Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives*. Springer, 2013, pp. 231–246.

[2] J. Zenkert and M. Fathi, "Multidimensional knowledge representation of text analytics results in knowledge bases," in *2016 IEEE International Conference on Electro Information Technology (EIT)*. IEEE, 2016, pp. 0541–0546.

[3] J. Zenkert, A. Klahold, and M. Fathi, "Knowledge discovery in multidimensional knowledge representation framework - an integrative approach for the visualization of text analytics results," *Iran Journal of Computer Science*, pp. 1–18, 2018.

[4] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from wikipedia," *Artificial Intelligence*, vol. 194, pp. 151–175, 2013.

[5] F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, and X. Rong, "Data mining for the internet of things: literature review and challenges," *International Journal of Distributed Sensor Networks*, vol. 11, no. 8, p. 431047, 2015.

[6] B. W. Weber and K. E. Niemeyer, "Chemked: A human-and machine-readable data standard for chemical kinetics experiments," *International Journal of Chemical Kinetics*, vol. 50, no. 3, pp. 135–148, 2018.

[7] M. Lehtonen, R. Petit, O. Heinonen, and G. Lindén, "A dynamic user interface for document assembly," in *Proceedings of the 2002 ACM symposium on Document engineering*. ACM, 2002, pp. 134–141.

[8] J. Zenkert, A. Klahold, and M. Fathi, "Towards extractive text summarization using multidimensional knowledge representation," in *2018 IEEE International Conference on Electro Information Technology (EIT)*. IEEE, 2018.

[9] M. Newman, *Networks: an introduction*. Oxford university press, 2010.