Reliability-Based Aggregation of Heterogeneous Knowledge to Assist Operators in Manufacturing

Richard Nordsieck*, Michael Heider†, Alwin Hoffmann* and Jörg Hähner†
*XITASO GmbH IT & Software Solutions, Augsburg, Germany
Email: {richard.nordsieck, alwin.hoffmann}@xitaso.com
†Organic Computing Group, University of Augsburg, Augsburg, Germany
Email: {michael.heider, joerg.haehner}@informatik.uni-augsburg.de

***Abstract*—In socio-technical systems, such as manufacturing processes, human operators are often entrusted with cognitive tasks that rely on tacit knowledge. Extracting the operators' tacit knowledge is beneficial to facilitate knowledge transfer and enable semantic machine learning. We improve upon an existing methodology that relies on operators' insights into influences on their decision making processes to extract tacit knowledge. By introducing a data-based weighting of the operators' information, we are able to control varying degrees of worker reliability and other individual biases, increasing the quality of the aggregated knowledge. We evaluate several methods to weigh and aggregate knowledge on a real-world dataset collected in the domain of fused deposition modelling (FDM) showing an improvement of 34% over a previously published baseline applied to our data. Applicability of the approach in the same domain is demonstrated by a case study, where the aggregated knowledge is utilised to shorten the time required for parametrisation.**

***Index Terms*—knowledge extraction, knowledge aggregation, collaborative knowledge capturing**

I. Introduction

In manufacturing scenarios, machines or production lines need to be parametrised to produce products of sufficient quality. To arrive at a parametrisation, operators perform an iterative cycle which involves quality assurance personnel inspecting the produced product's quality. Then, one or more process parameters which they deem suited to mitigate the observed quality defects are adjusted. This is a complex task requiring experienced personnel which highlights the need for improved knowledge transfer between operators. With regard to an increasing shortage of skilled operators, it is more and more important to collect existing knowledge of experienced operators and to make it available, e. g. by passive assistance systems.

In contrast to traditional knowledge extraction approaches, e. g. interviews [1], observations [2] or combinations thereof [3], data-based knowledge extraction has recently been proposed for manufacturing scenarios [4]. Here, observed parametrisation-quality tuples are aggregated in a knowledge graph with additional information regarding the experts' thought process, i. e. influences on their chosen parametrisation, which are collected with the human-machine-interface (HMI) during production. A knowledge graph from which a succinct rule base can be extracted, i. e. reliable information on how machine parametrisation influences a product's quality, is an important prerequisite for an operators' passive assistance system in production.

However, until now the approach did not generalise well to larger datasets and heterogeneous knowledge sources, i. e. operators with varying level of experience and cooperativeness. Naively aggregating sample-based data leads to erroneous relations in the knowledge graph, since the information the relations are based on could be erroneous or highlights non-causal influence–parameter change relationships. Reasons for erroneous information could be non-atomic influences, i. e. unclear cause-effect relations that can occur because multiple influences were highlighted for multiple parameters adapted, inexperienced operators, untrustworthy operators or operators that tend to have a higher trust in their experience than warranted [5].

In order to create a robust rule base for an operators' passive assistance system, there is the need to aggregate the sample-based data in a reliable and correct way. The main contribution of this paper is to show how semi-structured knowledge collected in production can be aggregated and validated to build such a robust rule base, serving as a less intrusive, data-based alternative to traditional knowledge extraction techniques. Based on a formal description of the domain, we introduce and evaluate several methods to weigh and filter the provided knowledge. Thereby, we improve the results obtained in [4] by 34%. Finally, we show how to automatically extract rules from the aggregated knowledge graph which are (1) quantified and (2) human-readable in natural language so that they can be directly used in a passive assistance system in case of quality defects.

The paper is structured as follows: Our case study on fused deposition modelling based additive manufacturing is introduced in Section II. In Section III, we present our methodology including the formalisation of the problem, heuristics to weigh and filter knowledge, and the extraction of quantified rules. An evaluation of our approach is given in Section IV and related work is presented in Section V. Whereas Section VI gives an outlook, Section VII concludes the paper.

II. Case study

In this study we extend the preliminary investigations performed by Nordsieck et al. [4] in the context of fused deposition modelling (FDM) based additive manufacturing [6]. Here,

an operator has to satisfy quality or other target criteria by executing multiple iterations of an iterative parameterisation process. Over the course of about two years, the operators in the case study performed a total of about 1200 iterations, of which 411 provide additional information (cf. Section III-A) about the *operator reasoning* on which this work is based. The iterations containing additional information were performed by nine operators of differing training and process experience levels.

The FDM process offers over 600 adjustable software parameters, although, according to our experience as well as the data we gathered, operators tend to only variate a small number of these. In our case study operators adjusted a total of 58 parameters while 34 of these were adjusted in more than 5% of iterations and 8 were adjusted in more than 25% of instances. Individual machines, as well as differing materials and, most importantly, object geometries impact what parameters need to be set, which leads to frequent reconfiguration for optimal production.

The 411 examples distribute onto 203 such parameterisation processes with the longest needing 14 manual readjustments of parameters, while 144 processes saw no reconfiguration after the initial parametrisation. When considering the 59 parametrisation processes with more than one iteration, the mean length was 4.5 with a standard deviation of 3.4 iterations. While we did provide a "best guess" default configuration (individually set for each printer) operators rarely relied on it for the initial parametrisation. Only 24 parametrisation processes started with the default, of which nine were immediately successful, while the others needed additional operator readjustments. This illustrates the great need for experienced operators and, until operators have gained this experience, the need to assist them in decision making for which we propose the following extensions on prior work.

III. Methodology

In this Section, we formalise the problem of extracting rules and present our approach to ascertaining the reliability of information provided by the operators and use it to influence the knowledge graph aggregation. Furthermore, we present a methodology to extract quantified human-readable rules in natural language which then form the basis for our real world evaluation.

A. Problem Formalisation

Based on Definitions 3.1 and 3.2 (previously presented in [4]), we establish the following preliminaries for this work. Non-adjustable factors $\mathcal{A} = \mathcal{O} \sqcup \mathcal{C} \sqcup \mathcal{T}$, i. e. object characteristics $\mathcal{O}$, target criteria $\mathcal{C}$ and environmental factors $\mathcal{T}$, influence a manufacturing process. Based on these, the operator chooses a parametrisation $p \in \mathcal{P}$ which is iteratively improved for each iteration $i \in \mathcal{I}$ of the production process until the target criteria $t \in \mathcal{C}$ are met:

Definition 3.1 (Parametrisation process): For an iteration $i+1$, the parametrisation $p_{i+1} \in \mathcal{P}$ is determined by

$$p_{i+1} = f(\{o, c, \tau\}, \{p_0, \dots, p_i\}, \{q_0, \dots, q_i\}, \mathcal{M})$$

Consequently, it is dependent on previous experience $\mathcal{M}$, non adjustable factors $o \in \mathcal{O}, c \in \mathcal{C}, \tau \in \mathcal{T}$, process parameters at previous iterations $p_0, \dots, p_i \in \mathcal{P}$ and previously achieved qualities $q_0, \dots, q_i \in \mathcal{Q}$. The parametrisation process is concluded by selecting the optimal parametrisation: $\hat{p}^* = \arg\max_p \{q_p\}$.

Upon choosing p_{i+1} the operator is given the opportunity to highlight their influences α_{i+1} in the machines' HMI along with their confidence $\zeta_{i+1} \in [0,\ 1]$ in the correctness of α_{i+1}.

Definition 3.2 (Influence): An influence α_{i+1} for the choice of a parameterisation p_{i+1} is given by:

$$\alpha_{i+1} = \{x \,|\, x \in \mathcal{A} \sqcup \mathcal{P} \sqcup \mathcal{Q} \text{ and } x \text{ influences the choice of values for } p_{i+1})\}$$

For a detailed definition of methods for knowledge graph aggregation refer to [4]. While we use the resulting knowledge graph as a data structure, it is not specifically relevant to the understanding of this work.

To illustrate the problem and its formalisation we consider the following simplified examples of two parametrisation processes as described in Section I: operators are tasked to manufacture two different objects *a* and *b* that have not been produced previously. Therefore, each of the objects constitutes a separate new task. They start with the first object *a* and use the default parametrisation $p_i \in \mathcal{P}$ with $p_i = (4, 5, 6)$, for the first iteration $i = 1$. Assume, this results in quality $q_i^a \in \mathcal{Q}$ with $q_i^a = (1, 0, 2)$. Note, that we measure quality defects in our case study, so a quality of 0 in each dimension would be considered optimal. To improve the third quality aspect $q_{i,3}^a = 2$, which then constitutes an influence α_{i+1} for the next iteration $i+1$, the operator adjusts the parametrisation to $p_{i+1}^a = (10, 5, 5)$, which results in $q_{i+1}^a = (1, 0, 0)$. This information can then be used to aggregate knowledge graphs [4] containing relations, which in turn can be used to extract rules.

Definition 3.3 (Relation): A relation $\eta_{x,y}$ is defined as the tuple $(q_{\cdot,x}, p_{\cdot,y})$ between a quality $q_{\cdot,x} \in Q$ and a parameter $p_{\cdot,y} \in P$.

Note that $q_{\cdot,x} \in Q$ selects the x^{th} element from $q_\cdot \in \mathcal{Q}$ of an unspecified iteration. This applies analogously to $p_{\cdot,y}$.

Definition 3.4 (Rule): A rule $r_{x,y}$ is defined as a tuple $r_{x,y} = (\eta_{x,y}, \nu) = (q_{\cdot,x}, p_{\cdot,y}, \nu)$ with a quality characteristic that will be improved upon $q_{\cdot,x} \in Q$ and a parameter $p_{\cdot,y} \in P$ that should be adjusted for the next iteration by a quantification $\nu \in \mathbb{R}$ with $\nu = \Delta(p_{i,y}, p_{i+1,y}) \to p_{i+1,y} - p_{i,y}$ describing the amount by which the parameter should be adjusted.
Based on the sign of the parameter change a conclusion of the rule can be defined as increasing if the quantification is positive and decreasing if the quantification is negative.

Revisiting the example, we could conclude a set containing two rules, $R_a = \{r_{3,1}^a = (q_{\cdot,3}, p_{\cdot,1}, 6), r_{3,3}^a = (q_{\cdot,3}, p_{\cdot,3}, -1)\}$, for the first task since it is not clear which specific parameter change affected the quality characteristic $q_{i+1,3}$. Note, that the parameter-quality relations are not atomic since an operator cannot highlight relationships between parameters and quality characteristics on an atomic level.

For the second task, i. e. manufacturing object b, the operators also start with the default parametrisation $p_i = (4, 5, 6)$. However, due to different object characteristics this leads to quality $q_i^b = (3, 1, 1)$. Now, they chose $p_{i+1}^b = (11, 5, 6)$, i. e. increasing $p_{i+1,1}^b$ by 7, which leads to quality $q_{i+1}^b = (0, 0, 0)$. The resulting set with rules is $R_b = \{r_{1,1}^b = (q_{\cdot,1}, p_{\cdot,1}, 7), r_{2,1}^b = (q_{\cdot,2}, p_{\cdot,1}, 7), r_{3,1}^b = (q_{\cdot,3}, p_{\cdot,1}, 7)\}$.

To be able to compare two rules we define three equality operators:

Definition 3.5 (High-level equality $=_h$): Influence and parameter have to concur for both rules:

$$r_{x,y} =_h r_{m,n} \iff x = m \wedge y = n$$

Definition 3.6 (Mid-level equality $=_m$): In addition to fulfilling $=_h$, both rules need to adjust the parameter in the same direction:

$$\begin{aligned} r_{x,y} =_m r_{m,n} \iff \quad & r_{x,y} =_h r_{m,n} \\ & \wedge \operatorname{sgn}(\nu_{r_{x,y}}) = \operatorname{sgn}(\nu_{r_{m,n}}) \end{aligned}$$

Definition 3.7 (Low-level equality $=_l$): Additionally to fulfilling $=_m$, the suggested parameter change needs to be similarly quantified, i. e. within 30%, in regards to range (cf. Section III-C) or absolute value:

$$\begin{aligned} r_{x,y} =_l r_{m,n} \iff \quad & r_{x,y} =_m r_{m,n} \\ & \wedge (\quad |\nu_{r_{m,n}}| * |\nu_{r_{x,y}}| \leq |0.3 * \nu_{r_{x,y}}| \\ & \quad \vee |\nu_{r_{m,n}}| * |\nu_{r_{x,y}}| \leq |0.3 * \nu_{r_{m,n}}|) \end{aligned}$$

Applied to the rules of our examples this leads to $r_{3,1}^a =_l r_{3,1}^b$.

The main problem this paper addresses is finding an aggregation function that yields a minimal set of correct rules

$$R_= = \{r | r \in Q \times P \times \mathbb{R}\}_=$$

regarding a given level of abstraction $= \in \{=_h, =_m, =_l\}$, of which parameters to apply to mitigate all quality defects that can occur in a given (re-)parametrisation process that conforms to the process described in Section I. In our example, a possible candidate for one such rule set could be $R_{=_l} = R^a \cap R^b$, where $\cap$ is defined for equality under $=_l$ and each ν in $R_{=_l}$ is calculated as the mean of the corresponding quantifications from R^a and R^b. To arrive at the optimal rule set, in practice, we utilise a knowledge graph as a data structure which is aggregated by the methods presented in Section III-B. The quality of the knowledge graph is measured by comparing extracted rules to an expert validated ground truth.

B. Ascertaining Reliability of Operator Provided Information

In practice, and illustrated by the example in Section III-A, operators adjust more than one process parameter at once leading to non-atomic insights, i. e. multiple influences that are in relation with multiple process parameters for example if they want to limit the amount of iterations needed during the parametrisation process. Also, process parameters and influences are multi-dimensional with over 600 process parameters and 14 quality characteristics. This high dimensional space can be reduced to 58 and 14, respectively by relying only on data that is present in influences α that highlight the relevant parameter & quality subset for an iteration. Apart from the high dimensionality, few examples, the unsupervised nature of the problem, as well as operator and environmental biases, illustrate that identifying the process parameter(s) influenced by a singular influence and vice versa is a non-trivial task that is best addressed with heuristics. Consequently, we design methods to ascertain the reliability of information provided by the operators with the goal of filtering out erroneous information. The underlying idea is to first weigh the relations and then filter them with an adaptive threshold.

We investigate methods that are applied on data contained in insights α before graph aggregation as well as methods that are applied after graph aggregation. For those that are applied before graph aggregation, we focus on operator confidence, frequency of the relation, as well as quality improvements. The *operators' confidence* ζ is directly contained in the insight of an iteration and does not need computation. *Relation frequency* is established by summing the occurrences of a relation over all iterations and dividing this by the number of unique relations. This is then assigned to the weight of the respective relations:

$$w_{F_{x,y}} = \frac{\sum_{i \in \mathcal{I}} (p_{i,x}, q_{i,y})}{|\{\eta_{x,y} \mid p_{\cdot,x} \in \mathcal{P}, q_{i,y} \in \mathcal{Q}\}|}$$

Quality improvement is given by calculating $\Delta(q_i, q_{i+1})$ or short Δq. These three techniques can be combined at will. Weights for relation frequency and quality improvement are multiplied if they are combined. Operator confidence, if combined, is factored in by $w/2 + w\zeta$, where ζ is the operators' confidence and w is the weight obtained by relation frequency F or quality improvement Δq, respectively.

Clustering-based weighting methods, denoted by C, are utilised to detect outliers and are applied to subsets of the iterations' attributes. Based on different combinations of subsets s_o with $o \in \{p, q, \alpha, \Delta p, \Delta q, \Delta\alpha\}$, OPTICS [7] is used for density-based clustering. We utilize the Minkowski distance and set the minimum amount of neighbours constituting a cluster to two. The resulting clusters are then used to weigh the relations of the knowledge graph by counting the number of iterations contributing to the respective relation that were successfully assigned to a cluster, thereby discounting the outlying iterations that were not successfully assigned a cluster. This is then normalised by the amount of iterations and used as the relation weight:

$$w_{C_{s,\eta}} = \frac{|\{i \in \mathcal{I}, i \in \text{OPTICS}(s), \eta(i)\}|}{\{i \in \mathcal{I}, \eta(i)\}},$$

where OPTICS generates sets of clustered iterations and unclustered iterations are not included in any sets and $\eta(i)$ determines if a specific relation $\eta_{x,y}$ can be extracted from the given iteration i.

Another clustering-based approach, denoted by CC is based on the assumption that similar parametrisations should lead to similar qualities. Therefore, iterations that are clustered in parameter space should also be clustered in quality space. We

prepared subsets s_o, with $o \in \{p, q, \Delta p, \Delta q\}$. The subsets are used to calculate weights analogous to the clustering methodology described above differentiating between relative and absolute quality and parameters:

$$w_{CC_{s,\eta}} = \frac{|\{i \in \mathcal{I}, i \in \text{OPTICS}(s_p) \cup \text{OPTICS}(s_q), \eta(i)\}|}{\{i \in \mathcal{I}, \eta(i)\}}$$

In contrast to the previous methods, the *influence validity* method, denoted by IV, directly tries to quantify the correctness of the operator's assumption that the changed process parameters are improving the highlighted quality characteristics. To achieve this, the sum of the highlighted quality characteristics is divided by the overall quality for each iteration contributing to the relation. The resulting fractions are then summed and normed by the amount of iterations that contributed to this relation and used as its weight:

$$w_{IV_\eta} = \frac{\sum_{\{i \in \mathcal{I}, \eta(i)\}} \frac{\sum_y q_{i,y}}{\sum_y q_{i+1,y}}}{|\{i \in \mathcal{I}, \eta(i)\}|}$$

The adaptive threshold is computed by taking the mean $\bar{x}$, median $\tilde{x}$ or elements smaller than the first quartile Q_1 over all relations. Based on this, relations whose weight is smaller than the threshold are removed from the knowledge graph.

C. Extraction of Quantified Rules

To extract quantified rules, we rely on the aggregated knowledge graph to identify high-level relations between influences (i. e. quality characteristics or environmental) and adapted parameters. Given the quantified rules rely on the aggregated knowledge they represent aggregations of iterations and are therefore independent of decisions at a specific point in time. In the following we will rely on these aggregated rules. Quantification of aggregations is non trivial, however, by analysing p and Δp over all iterations contributing to the relation that exhibited the respective high-level relation, we are able to discern both the conclusion (i. e. increase, decrease, set) as well as the parameter quantification of the conclusion of the rules. A parameter can be quantified in three ways: (1) range, (2) step size and (3) concrete values (usually used for categorical parameters). Range is quantified by $\mathbb{E}(p_{\cdot,x}) \pm \sigma(p_{\cdot,x})$. Step size is quantified by $\mathbb{E}(\Delta p_{\cdot,x})$. Concrete values in the case of categorical parameters are quantified by the most common value.

IV. Evaluation

To evaluate our approach, we apply the methods presented in Section III to data collected in a case study with FDM described in Section II. We determine a ground truth, benchmark different aggregation methods on the ground truth, evaluate requirements in regards to sample size and analyse the real-world applicability of utilising the collected and aggregated knowledge in a proof-of-concept case study.

A. Ground Truth Creation

Well defined ground truths are not available for the FDM domain. Also, unstructured knowledge bases e. g. Simplify3D's troubleshooting guide, from which they could be created are unsuited since the methodology underlying their creation is unknown. Therefore, we describe a methodology to obtain a ground truth against which the baseline of [4] as well as the results of our approach can be compared.

Three FDM experts are independently tasked with classifying whether the rules generated by the baseline are correct on a high level of abstraction, i. e. whether there exists a relation between parameter and condition. If that is the case they proceed to evaluate the rule at a medium and, lastly, a low level. As described in Section III-A, medium equality is achieved if the rules' action, e. g. increase or decrease, is suitable, whereas low equality is achieved if the quantification—the amount the parameter should be adjusted—is within 30% of the experts' opinion. If a rule is unequal at low or medium level the experts' adjust it according to their knowledge.

The resulting individual rule sets are merged by averaging—mean for numerical, most frequent value for categorical—quantifications and assigning an action accordingly if at least two experts validated a rule for a given relation. The now aggregated rule set is suitable for a use as ground truth since multiple experts' evaluated the given rules in a detailed manner, providing corrections where necessary. Especially, due to the experts' corrections the ground truth is not a subset of the baseline which would aggravate the comparison between baseline and aggregation method. Still, it is likely to focus on a certain subset of the experts' knowledge. However, since this subset is informed by parameter choices encountered in practice it could be argued that it focusses on the most practically relevant area of expertise. The experts provided a mean of 41 rules each with a standard deviation of 2.16 rules. All experts agreed on 20 rules, 21 different rules were confirmed by two experts, whereas another 21 rules were only approved by a single expert. This highlights the individuality of experts that naturally gained their experience on different printers, materials and objects and highlights the importance of our approach to only accept rules that are agreed upon by at least two experts. As the number of involved experts and the fractural nature of their knowledge is similar to industrial settings, we assume that the process of ground truth creation is applicable to other industrial domains. In practice, the ground truth creation process could also be used as an editorial step to increase the quality of knowledge extracted with our approach.

B. Evaluation Against Ground Truth

To benchmark our proposed methodology and weighting methods, we compare them against the ground truth generated with the methodology presented in Section IV-A.

To evaluate overlap between rules contained in the aggregated knowledge graph and the ground truth, we utilize the three equality operators defined in Section III-A, which are suited to evaluate rules on differing degrees of abstraction. To be able to evaluate metrics on rule sets, e. g. ground truth

TABLE I
RESULTS FOR SELECTED AGGREGATION METHODS COMPARED TO THE GROUND TRUTH (SEE SECTION IV-B). BEST PERFORMING METHODS (ACCORDING TO F1 SCORE) PER LEVEL ARE HIGHLIGHTED IN BOLD. FOR THE PATTERN UNDERLYING THE COMPOSITE METHOD NAMES REFER TO SECTION III-B. AS BASELINE WE APPLIED THE METHODOLOGY PRESENTED BY NORDSIECK ET AL. [4] TO OUR DATA.

Level	Method	Precision	Recall	F1	# rules
high	baseline	0.34	1.00	0.51	121
	$\zeta\, F\, \Delta q \;\#\; \bar{x}$	**0.74**	**0.61**	**0.67**	**34**
	$C\, p\, q \;\#\; \bar{x}$	0.55	0.83	0.66	62
	$IV\, \alpha\, q \;\#\; Q_1$	0.42	0.95	0.59	92
	$CC\, p\, q \;\#\; Q_1$	0.40	0.71	0.51	73
mid	baseline	0.26	0.76	0.38	121
	$\zeta\, F\, \Delta q \;\#\; \bar{x}$	0.41	0.34	0.37	34
	$C\, p\, q \;\#\; \bar{x}$	**0.40**	**0.61**	**0.49**	**62**
	$IV\, \alpha\, q \;\#\; Q_1$	0.32	0.71	0.44	92
	$CC\, p\, q \;\#\; Q_1$	0.27	0.49	0.35	73
low	baseline	0.17	0.49	0.25	121
	$\zeta\, F\, \Delta q \;\#\; \bar{x}$	0.06	0.05	0.05	34
	$C\, p\, q \;\#\; \bar{x}$	**0.27**	**0.41**	**0.33**	**62**
	$IV\, \alpha\, q \;\#\; Q_1$	0.20	0.44	0.27	92
	$CC\, p\, q \;\#\; Q_1$	0.16	0.29	0.21	73

and prediction, of differing size and ordering an evaluation set is needed. An evaluation set of two rule sets is created by comparing each given rule with the given equality operator. If the rules are equal according to the equality operator, a substitution is carried out, if it is unequal it is treated as a unique rule. Substituted rules, as well as unique rules of both rule sets are concatenated to form the evaluation set. Referring back to the example from Section III-A, $R^{a,b} = \{r^{a,b}_{1,1}, r^{a}_{1,3}, r^{b}_{2,1}, r^{b}_{3,1}\}_{=_l}$ would be the evaluation set for R^a and R^b for $=_l$. Both rule sets are compared against the evaluation set resulting in binary sets that can be directly compared. We evaluate the similarity between a prediction and ground truth with prediction, recall and F1 score (values in [0,1], higher is better). Also, the number of rules resulting from using a given aggregation method is reported.

We apply the aggregation methods described in Section III-B and evaluate them according to these metrics for the three levels of abstraction. Note that the low abstraction level is the one with the greatest relevance for industrial use cases since it provides data-based quantifications, which are often not available using traditional knowledge extraction methods. Table I shows the baseline achieved by applying the methodology of [4] to our data, the methods that performed best (in regards to F1 score) on high-, mid- and low-level of abstraction as well as the best performing representative of each class of methods.

Firstly, we can observe that precision, recall and F1 increase for all methods with increasing level of abstraction. This is intuitively explainable by considering that it is much harder to forecast the exact parameter quantifications compared to conclusion or even the pure existence of relations. The number of rules stays constant since the level of abstraction used during evaluation is independent of the aggregation method.

Some methods, e. g. $\zeta\, F\, \Delta q \;\#\; \bar{x}$, seem to be too excessive in filtering suspected erroneous information. The fact that there is a comparatively very sharp decrease of precision, recall and F1 with decreasing level of abstraction can be an indicator of that. Apparently, even though very good results are achieved on a high level of abstraction with a comparatively small number of rules, data points necessary for correctly calculating the quantifications are discarded. Also, very low rule counts can be an indicator of too aggressive filtering. Likewise, methods relying on relative changes in parametrisation or quality perform worse than those which rely on absolute values. One possible explanation for this behaviour is that the corresponding data scarcity limits the performance of the clustering algorithm. This is bolstered by the observation that OPTICS detects fewer large clusters on relative than on absolute data. Compared to clustering based approaches, $IV\, \alpha\, q \;\#\; Q_1$ retained a relatively large number of rules and therefore obtains high recall results while also moderately increasing precision. Therefore it seems like it is a candidate for further development that could thrive with a more aggressive filter method. Clustering in both parameter and quality spaces separately (CC), however, seems to be an unfeasible approach for outlier detection since relatively many rules are retained while suffering in recall and precision. A possible explanation could be that individually clustering the respective spaces leads to fewer clusters found than clustering on the combined p-q space as evidenced by the good performance of $C\, p\, q \;\#\; \bar{x}$ which is based on the same underlying data, albeit in a different format.

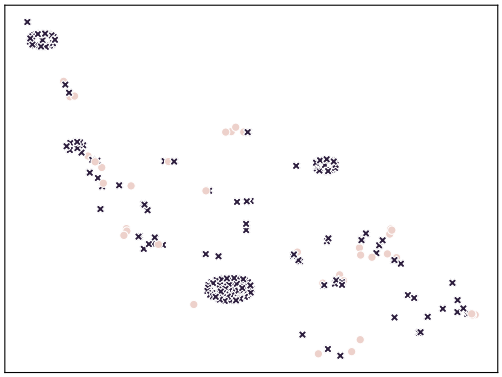

Fig. 1. t-SNE visualisation of quality space q. Elements that were assigned a cluster with OPTICS are highlighted by a black "x" , those that were unclusterable for OPTICS are represented as grey dots.

To get a qualitative understanding of OPTICS' clustering quality, we used t-SNE [8] as a dimensionality reduction approach to show all available iterations in two dimensions and highlighted those iterations that were successfully clustered by OPTICS. Figure 1 visualises the clustering obtained in quality space q. Here, we can observe that elements within large neighbourhoods as visualised by t-SNE are also detected as

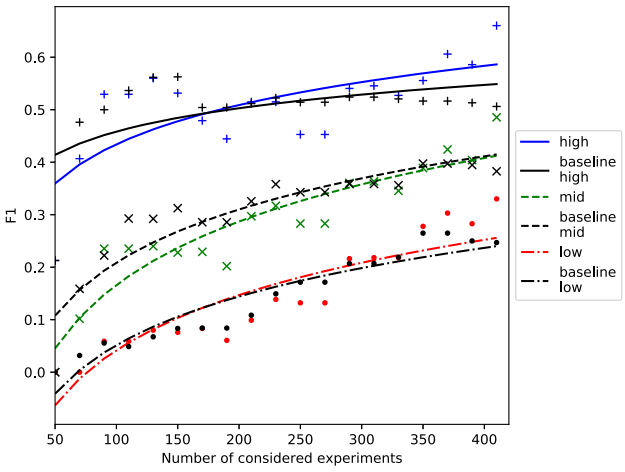

Fig. 2. F1 score for increasing number of considered samples of the aggregation method $C\,p\,q\,\#\,\bar{x}$ and baseline for high, mid and low abstraction levels.

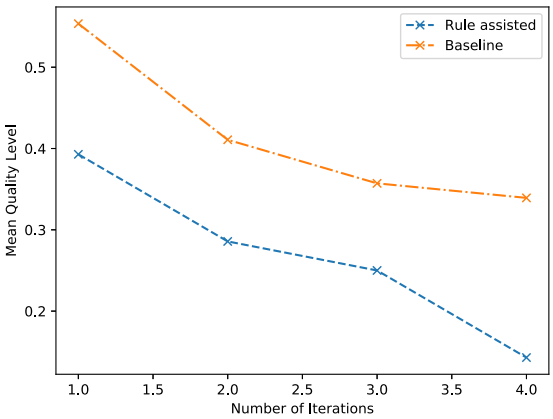

Fig. 3. Mean quality level (lower is better) per iterations as achieved by the baseline as well as rule assisted group. For comparability the figure is truncated at iteration 4.

belonging to clusters by OPTICS. While some iterations that appear outliers in the t-SNE representation are still clustered by OPTICS, we conclude that overall it is capable to discern the respective clusters and function as a suitable outlier detector.

In general, recall seems to be higher than precision, with the exception of $\zeta\,F\,\Delta q\,\#\,\bar{x}$. The baseline outperforms our methods on recall, however, lacks in precision. Consequently, it is outperformed by a margin of 0.16, 0.10 and 0.08 on high, mid and low abstraction levels which translates to 32%, 27% and 34%, respectively on the F1 score by method $\zeta\,F\,\Delta q\,\#\,\bar{x}$ for the high abstraction level and $C\,p\,q\,\#\,\bar{x}$ otherwise. These results are achieved while reducing the amount of rules by 49%, which in turn facilitates their use in a passive assistance system since the rule application is, intuitively, easier if there are fewer rules to choose from.

Overall, we have to conclude that there does not seem to be a single method that is best suited to all levels of abstraction. However, since only the lowest abstraction level is relevant in practice and $C\,p\,q\,\#\,\bar{x}$ provides consistently good performance through all abstraction levels which is indicative of good generalization characteristics, we propose it as the best method, which will be used in the following experiments.

C. Samplesize Requirements

To ascertain the effect of different amounts of sample data, i. e. available process iterations, on the aggregated knowledge, we evaluate the method that performed best in Section IV-B on the expert validated ground truth as measured by the F1 score. Hence, both baseline and $C\,p\,q\,\#\,\bar{x}$ are evaluated with increasing sample sizes, which are included in the order of their collection, in the range of 50 to 410 in steps of 20. We assume that, similar to learning systems, the performance converges to an optimum after a specific amount of time. To visualise whether convergence is reached, logarithmic curves are fitted to the predictions. The results are shown in Figure 2.

Generally, increasing amounts of samples positively influence the predictive performance. However, this effect is not linear or at least heavily noised as is evidenced by the drop in performance for 170, 190 and 210 examples, respectively. Possible causes for the drop could have been an explorative operator behaviour, e. g. of an inexperienced operator or because of changes in material. The fact that this drop is especially evident in the higher abstraction levels is probably due to the insufficient performance on the low abstraction level for low sample sizes.

While our initial assumption of a convergence seems to be true for the baseline since there is no notable improvement for the last 20% of the sample, and even more on the high abstraction level, the same cannot be said for our aggregation method. $C\,p\,q\,\#\,\bar{x}$—apart from the aforementioned drop around 200 samples—is gradually increasing for the first 300 examples. At the last 100 examples an increase with a significantly higher gradient is visible. This could be caused by relations being already present in the baseline since they were encountered before in iterations classified as outliers by the aggregation. This explanation could be aided by the observation that the difference between results of the baseline for mid and high abstraction levels to low and mid abstraction levels of the aggregation method is decreasing with increasing number of examples, almost reaching equality at 410 examples, hinting at the effect of sufficient data to calculate meaningful quantifications and conclusions.

Overall, it is evident that a sample of 410 is insufficient to reach a plateau. Still, this technique can be used to judge whether data collection should continue or whether a plateau of the extracted rule set is reached.

D. Applicability in Practice

To study the applicability of the filtered aggregated rule set in practice, e. g. as a passive assistance system, we utilised the

case study described in Section II. We selected six participants of varying experience, that did not previously participate in data collection or ground truth creation. They were presented with a printing task designed to challenge several quality characteristics. The target criteria was to optimize both quality and iterations required. To ensure a comparable environment, each participant utilized the same printer with the same material and the same initial parametrisation.

The participants were divided in two groups of three. The baseline was created by the control group that completed the task solely relying on their previous experience and process knowledge. The second group was given the rule set aggregated by $C\,p\,q\,\#\,\bar{x}$ (cf. Rule 4.1 for an example rule).

Rule 4.1: If you encounter warping, try to incrementally increase the parameter material_bed_temperature by 5.2.

To conform with the human factor to be expected in production scenarios, the operators in the second group could apply as many rules pertaining to observed quality defects as they saw fit. However, their degree of FDM relevant expertise was significantly lower than that of the baseline group. The uneven split regarding operator experience was chosen to evaluate to which degree novices (with rules) are able to compete with experienced operators (without rules). The control group took a mean of 5.33 iterations with a standard deviation of 0.47, whereas the rule assisted group took an average of 5.0 iteration with a standard deviation of 1.4. The high standard deviation in the rule assisted group is explained by one operator that required 7 iterations compared to the 4 iterations required by the other operators.

A comparison of the achieved quality can be seen in Figure 3, in which the achieved mean quality (consisting of 4 quality characteristics) of both groups is compared. The figure is truncated at 4 iterations since otherwise the comparability is not ensured because of the different numbers of iterations required by the participants. While the environmental temperature was higher for the rule assisted group than the control group, operators should be able to compensate environmental temperature based on their experience. Therefore, this does not explain the general offset in performance. After iteration 4, the achieved quality of the baseline improved while the one operator of the group utilising rules decreases in achieved quality until finding a suitable parametrisation at iteration 7.

For eight of the observed 13 occurring quality defects, rules could be extracted. However, for some quality characteristics they were erroneous which lead all participants in the rule assisted group to be unable to significantly improve upon these quality characteristics. All the more interesting is the fact that for the first 4 iterations the application of the rules achieves a better quality faster than the control group, which consisted of more experienced operators. Apart from the general offset, this is underpinned by the fact that the difference between achieved mean quality at iteration 4 is larger than that at iteration 1. It can be noted that at least some operators had difficulties in discerning when the optimal attainable quality was reached, since they continued to explore parametrisations that did not improve upon the best quality they achieved. Also, in the rule assisted case different rules were applied. Due to the inexperienced operators, we assume that this hints at a high uncertainty while selecting rules. Improving this could further increase the applicability of the extracted rules in practice. Overall, the applicability of the extracted rules in practice can be assessed as positive since they lead to quicker parametrisations which attain a better quality faster than the control group. However, because of the small study size and environmental factors which are hard to control further investigations should be conducted.

V. Related Work

Our approach is directly related to data-based knowledge extraction, significantly improving the knowledge aggregation methodology described by Nordsieck et al. [4].

Also, it relates to the problem of knowledge graph completion, as candidate triples have to be checked whether to be included in the knowledge graph or not. Consequently, work done on filtering candidate triples by Borrego et al. [9] is conceptually related. However, the main difference is that in knowledge graph completion candidate triples are usually generated based on information contained in the knowledge graph, while in our case they are provided by process data and operator information gained in a manufacturing process. Also, in KG completion an existing graph is further refined whereas in our case the initial graph has to be created, which leads to inherent differences in filtering methodology.

The field of Organic Computing (OC), dealing with complex heterogeneous systems similar to our socio-technical system use case, defines the multifaceted concept of *trust* [10]. One important aspect of which is *credibility*, that assesses good faith participation and competence of partners. While we currently handle this issue implicitly during aggregation, dedicated methods to establish trust might improve future results. Our scenario can be classified as a socio-technical self-adaptive system, with the operator adapting to observations of the manufacturing process. According to Ramirez et al. [11] the most relevant sources of run-time uncertainties are related to interactions of the system and its context. To address these, the application of subjective logic to aggregate run-time observations into actionable knowledge has been proposed by Petrovska et al. [12]. However, since their methodology is centred on cyber-physical systems and ours on human operators it is not directly transferable.

Discovering *causal relationships* based on observations [13] is another active research topic that could be utilised to ascertain reliability of operator given information. Since deep learning based approaches usually require large datasets, we assume that the limited and sparse data available in our scenario would hamper its impact.

Clustering over (explainable) knowledge graph embeddings [14], [15] could also serve a similar purpose. However, the limited sample size available in our scenario probably is limiting the achievable quality of embeddings. Also, semantic representation capabilities of embeddings have recently been questioned [16].

VI. Outlook

Even though the presented methods are an improvement to the unfiltered baseline, the aggregation mechanism will be further refined. Classifying whether an operator is building knowledge through exploring behaviour or exploiting knowledge in a given sample could be addressed by novelty detection [17], which could then be used to discount information gathered during explorative behaviour. Alternatively, clustering of knowledge graph embeddings [15] could be investigated if a suitable representation for actual parametrisations is found. To achieve this, however, a different representation of the knowledge graph is needed.

Furthermore, a detailed evaluation of rules, focussing on differences between quality characteristics or parameters could point towards uncertainties of the operators. We plan to repeat the experiment designed to determine applicability in practice (cf. Section IV-D) on a larger scale to gain more dependable results. Since operators exhibited difficulties in selecting rules to apply we plan to research ways towards a better operator guidance. One approach would be limiting the amount of shown rules to those that promise the best results. Another would be the quantification of the preliminaries of rules, e. g. quantifying quality, which could lead to rules with a better defined scope that limits the amount of rules that could be applied. Also, extending the rule definition to handle multiple preliminaries would add the ability to correctly include environmental factors such as temperature and material as additional preliminaries which would narrow their applicability.

Another approach towards greater applicability in practice is an improvement of the aggregation methods. Candidates that we plan to explore are combinations of aggregation methods in ensembles and a stronger focus on explicit atomisation of p-q relations. We will determine whether the ordering of the input introduces bias and investigate whether shuffling will mitigate it. Also, different metrics for evaluating the methods' performance could be improved such as investigating what amount of the data is covered by which rules.

VII. Conclusion

In this paper, we presented the theoretical framework and an approach to improve the aggregation of data-based knowledge provided by operators of manufacturing processes during production by ascertaining its reliability with several weighing methods and filtering accordingly. In addition, a methodology to transform the knowledge contained within the knowledge graph to human readable rules has been proposed. The methods are evaluated against a ground truth created by experts, showing a clear best method with an improvement over the application of a previously published baseline to our data of 27% and 34% for the two lowest abstraction levels, which are especially relevant in industrial scenarios. Furthermore, the effect of sample size on the approach and thereby data-based knowledge extraction has been investigated. This analysis also provides benefits in practice, where it can be used as an indicator of rule set completeness, biases and previously undetected influences. Finally, the applicability of the quantified rules as a passive assistance system has been evaluated during manufacturing, showing promising preliminary results (successful parametrisation is found faster and is better than the one found by the control group at this iteration). Consequently, expanding and extending our approach may contribute to mitigate challenges arising from knowledge loss in manufacturing and support operators in complex (re-)parametrisations in the future.

References

[1] V. Deslandres and H. Pierreval, "Knowledge acquisition issues in the design of decision support systems in quality control," *European journal of operational research*, vol. 103, no. 2, pp. 296–311, 1997.

[2] N. Shadbolt, P. R. Smart, Wilson, and S. Sharples, "Knowledge elicitation," *Evaluation of human work*, pp. 163–200, 2015.

[3] L. Hörner, M. Schamberger, and F. Bodendorf, "Externalisierung von prozess-spezifischem mitarbeiterwissen im produktionsumfeld," *Zeitschrift für wirtschaftlichen Fabrikbetrieb*, vol. 115, no. 6, pp. 413–417, 2020.

[4] R. Nordsieck, M. Heider, A. Winschel, and J. Hähner, "Knowledge extraction via decentralized knowledge graph aggregation," in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*. IEEE, 2021, pp. 92–99.

[5] J. Kruger and D. Dunning, "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1121, 1999.

[6] I. Gibson, D. Rosen, and B. Stucker, *Additive Manufacturing Technologies - 3D Printing, Rapid Prototyping, and Direct Digital Manufacturing*. Springer-Verlag, 2015.

[7] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," *ACM SIGMOD Record*, vol. 28, no. 2, pp. 49–60, 1999.

[8] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[9] A. Borrego, D. Ayala, I. Hernández, C. R. Rivero, and D. Ruiz, "Generating rules to filter candidate triples for their correctness checking by knowledge graph completion techniques," in *Proceedings of the 10th International Conference on Knowledge Capture*, 2019, pp. 115–122.

[10] J.-P. Steghöfer, R. Kiefhaber, K. Leichtenstern, Y. Bernard, L. Klejnowski, W. Reif, T. Ungerer, E. André, J. Hähner, and C. Müller-Schloer, "Trustworthy organic computing systems: Challenges and perspectives," in *Autonomic and Trusted Computing*, B. Xie, J. Branke, S. M. Sadjadi, D. Zhang, and X. Zhou, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 62–76.

[11] A. J. Ramirez, A. C. Jensen, and B. H. Cheng, "A taxonomy of uncertainty for dynamically adaptive systems," in *2012 7th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*. IEEE, 2012, pp. 99–108.

[12] A. Petrovska, S. Quijano, I. Gerostathopoulos, and A. Pretschner, "Knowledge aggregation with subjective logic in multi-agent self-adaptive cyber-physical systems," in *Proceedings of the IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, 2020, pp. 149–155.

[13] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf, "Distinguishing cause from effect using observational data: methods and benchmarks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1103–1204, 2016.

[14] C. Wang, S. Pan, R. Hu, G. Long, J. Jiang, and C. Zhang, "Attributed graph clustering: A deep attentional embedding approach," *arXiv preprint arXiv:1906.06532*, 2019.

[15] M. H. Gad-Elrab, D. Stepanova, T.-K. Tran, H. Adel, and G. Weikum, "Excut: Explainable embedding-based clustering over knowledge graphs," in *International Semantic Web Conference*, 2020, pp. 218–237.

[16] N. Jain, J.-C. Kalo, W.-T. Balke, and R. Krestel, "Do embeddings actually capture knowledge graph semantics?" in *European Semantic Web Conference*. Springer, 2021, pp. 143–159.

[17] C. Gruhl and B. Sick, "Novelty detection with candies: a holistic technique based on probabilistic models," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 6, pp. 927–945, 2018.