Towards A Statistic Ontology for Data Analysis in Smart Manufacturing

Zhuoxun Zheng[1,2,*], Baifan Zhou[3], Dongzhuoran Zhou[3], Akif Quddus Khan[4], Ahmet Soylu[2] and Evgeny Kharlamov[1,3]

[1]Bosch Center for Artificial Intelligence, Germany
[2]Department of Computer Science, Oslo Metropolitan University, Norway
[3]SIRIUS Centre, University of Oslo, Norway
[4]Norwegian University of Science and Technology

Abstract

Statistical analytics has been playing an important role for uncovering patterns and trends from data in smart manufacturing. However, how statistical analytics are performed is usually stored in the underlying code and suffers from transparency and re-usability, which is vital for modern industry. To this end, we propose a statistical ontology *StatsOnto* that models not only the concepts in statistics, but also allows encoding the procedure of statistical analytics, which is limitedly addressed in past works. We present a preliminary evaluation of *StatsOnto* with a Bosch use case with a user study, competence question and coverage discussion in this poster paper.

Keywords

Ontology Engineering, Knowledge Graph, KG Generation, Data Science, Manufacturing

1. Introduction

Smart manufacturing is a term generally applied to improve manufacturing operations through system integration, linking of physical and cyber capabilities, and taking advantage of information including leveraging the big data analysis [1, 2]. In this process, statistical analyses have always played a crucial role, as they not only identify patterns and trends from large amounts of data [3], but can also be further used in methods such as machine learning [4, 5].

However, how statistical analytics are performed is usually stored in underlying code and suffers from transparency and re-usability, which is vital for modern industry [6]. Semantic technologies including ontologies are beneficial for improving the transparency since they offer a standardised way describing statistical analysis knowledge and procedure in machine processable formalisation that opens the door for many applications [7, 8], such as automated reasoning, optimisation of statistical pipelines [9, 10]. Currently there are a few studies that discuss partially statistical analytics pipeline modelling. For instance, the computer science ontology [11, 12] contains the general knowledge about statistics, but the concepts of specific statistical calculation are not involved. Statistics ontology [13, 14] enumerates the various statistical methods, but they insufficiently study the procedures of the statistical pipelines.

Hangzhou'22: The 21st International Semantic Web Conference, October 23–27, 2022, Hangzhou, China
*Corresponding author.
zhuoxun.zheng@de.bosch.com (Z. Zheng)

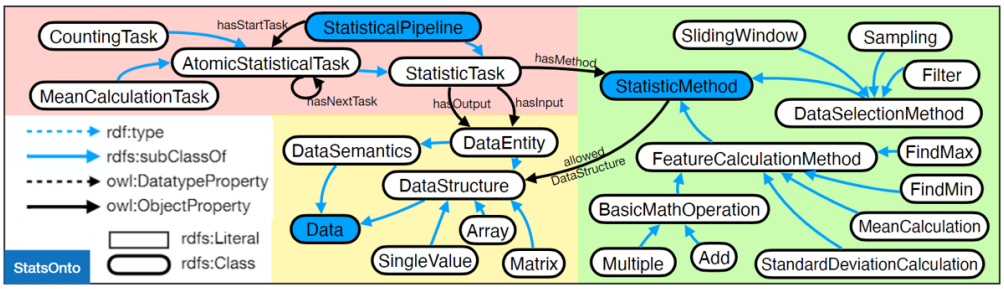

Figure 1: Schematic illustration of *StatsOnto* (partial). Three important classes *StatisticalPipeline*, *Data* and *StatisticalMethod* are colored in blue. Red, yellow and green backgrounds denote concepts in *Task*, *Data* and *Method*.

To this end, we propose a statistical ontology *StatsOnto* with the angle of procedure orientation. We discuss *StatsOnto* under an industrial scenario of smart manufacturing and use real-world data provided by Bosch [15]. On the one hand, *StatsOnto* covers statistical methods such as feature calculation (e.g., mean), sampling and filtering; on the other hand, *StatsOnto* also offers a set of classes for construction of statistical analytical pipelines, including statistical pipeline, various tasks, data entities, etc. This poster paper presents a preliminary evaluation of *StatsOnto* in the industrial scenario with user study, competence question and discussion of knowledge and scenario coverage.

2. Our Approach

Industrial Scenario. We limit our scope in this poster paper to an industrial scenario of smart welding manufacturing at Bosch [16], to show the real-world impact as well as examples. *StatsOnto* aims to help the engineers at Bosch to gain insights from the data collected from the welding production, and to monitor quality of the welding operations.

Requirements. We derive the requirements for *StatsOnto* based on the scenario purposes and discussion with users [17, 18, 19]: *R1. Procedure-Orientation:* *StatsOnto* should be able to reflect the statistical analytics procedure, allowing to describe sequence of statistical tasks in a data pipeline. This opens the door of knowledge graph based verification, reasoning, optimisation of statistical analytical pipelines. *R2. Transparency:* *StatsOnto* should improve the transparency of the representing statistical analytics in industry among engineers. *R3. Knowledge Coverage:* *StatsOnto* should cover the knowledge and practice of the statistical analytics, such as statistical method, data structure. *R4. Purpose Coverage:* *StatsOnto* should cover the four types of task: *data inspection* (e.g. find the data with certain property), *statistical modelling* (e.g. build the distribution of the data), *data denoising* (e.g. detect and remove the outliers) and *data analysis* (e.g. interpolation, subsampling).

Ontology Engineering Process. We broadly follow the routine of Ontology development 101 [20, 21], which is a kind of collaborative ontology engineering methodology. The whole process can be divided into the following 4 steps. *Step 1: Domain Analysis*, where common statistical analytics at Bosch are discussed. Common and important terms of statistical tasks are enumerated and classified. *Step 2: Concepts Formalisation*, where enumerated basic concepts are formalised as classes and relationships between them. *Step 3: Mechanism Investigation*, where the mechanism of how *StatsOnto* can serve as the basis in generating KGs which represent concrete statistical analystic pipelines. This step reflects the requirement of *Procedure-Orientation* of *StatsOnto*. *Step 4: System Deployment*, where *StatsOnto* will be deployed in manufacturing and

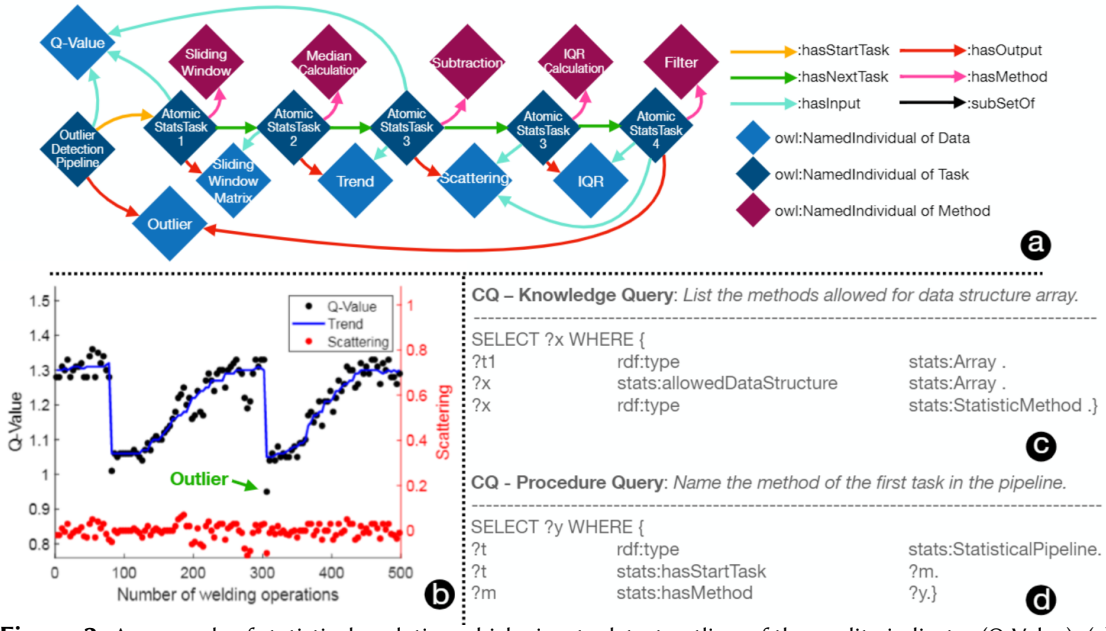

Figure 2: An example of statistical analytics, which aims to detect outliers of the quality indicator (Q-Value). (a) Schematical illustration of a KG generated based on *StatsOnto*, which represents a data pipeline for *data denoising* in the case study (R1-R2). (b) The visualisation of the task results. (c) Example Competence Questions (CQ) for knowledge query (R2-R3), and (d) for procedure query (R1-R2).

user feed-backs are collected constantly for iterative processing and further improvement.

Statistical Ontology is expressed in OWL 2 EL language for its expressivity and it is still polynomial for query answering [22]. It has 88 classes, 30 properties, and 484 axioms. In particular, there are three most important classes in *StatsOnto* (Fig. 1), *StatisticalTask*, *StatisticalMethod* and *Data*. *StatisticalTask* can be divided into two sub-classes, the *AtomicStatisticalTask* and *StatisticalPipeline*. The former models the basic and common statistical tasks, such as mean calculation task, counting task, etc., and connects to *StatisticalMethod* with property *hasMethod*. While *StatisticalPipeline* can be regarded as the serialization of *AtomicStatisticalTask*. Each individual of *StatisticalMethod* is a piece of script in computer language such as Python. There are also two kinds (sub-classes) of *StatisticalMethod*s, namely *DataSelectionMethod* and *FeatureCalculationMethod*, which correspond to the two basic steps in statistical analytics, namely *determination the target data of interest* and *calculation of the desired feature* respectively. Besides these concepts, *StatsOnto* also specifies some rules, which constraint the inputs or outputs of each *AtomicStatisticalTask* which invokes certain *StatisticalMethod*. For example, the following axiom specifies, that any task which has *Array* as input and has *MeanCalculationMethod* as method, has *SingleValue* as output: $\exists hasOutput^{-}.(Task \sqcap \exists hasInput.Array \sqcap \exists hasMethod.MeanCalculation) \sqsubseteq SingleValue$.

3. Evaluation, Conclusion and Outlook

User Study. We organised a workshop at Bosch and collected 28 reports from experts of different background, such as welding engineers, data scientists, knowledge engineers. We have also collected some typical statistical analysis tasks in Bosch's welding manufacturing (Tab 1). The users we divided into two groups; each group first perform a statistical analytical task without our method, and then answer several single-selected questions (SSQ) (Tab 1); after that, the two groups exchange their statistical tasks and do the tasks with our method, and then

Table 1
Example tasks and their description and examples of single selection questions (SSQ)

Tasks	Description
StatsTask1	Extract four statistics from a sequence: mean, std., min. and max.
StatsTask2	Compute the trend, scattering and outliers of a sequence with median filter, etc.
Questions (Q) and Answers (A) for SSQ	
Q1: What's the structure of the input data we used for the statistical analytics? A1: (A) Single features (B) Array (C) Matrix	
Q2: What method have we used in the task? I: median filter, II: mean filter, III: Gaussian sampling. A2: (A) I (B) I + II (C) II (D) II + III	

answer the SSQs. Fig. 2 demonstrate *StatsTask2*, which aims to detect the welding operations with abnormal quality indicator (Q-Value).

Transparency. After evaluation, the correctness of SSQ for the participants with the help of *StatsOnto*, no matter in which group, reaches 96.7%, while the correctness without *StatsOnto* is about 93.3%. The results show that the ontology indeed helps the users to better understand the statistical analysis pipeline, thus increase the transparency.

Case Study and Procedure Orientation. We use an example for *data denoising* (Fig. 2) for demonstrating the procedure orientation. Given the input of Q-Value *Array*, the pipeline first extract its trend by calculating the median value with sliding window, then calculates the scattering by the difference between the trend and Q-Value. The points with large scattering (large deviation from the trend) are detected as the outliers. This shows *StatsOnto* is capable of representing the procedure of statistical analytics.

Knowledge Coverage. We select two example CQs in SPARQL from two aspects: knowledge query (Fig. 2c), analysis procedure query (Fig. 2d). Results show that all these CQs return desired answers on KGs generated based on *StatsOnto* with the welding manufacturing data, demonstrating good *Knowledge Coverage* of *StatsOnto*.

Purpose Coverage. After extensive discussion in the workshop, we categorised most statistical analytical tasks in our project into the four types of purposes (R4): data inspection, statistical modelling, data denoising and data analysis, and found most of the purpose can be covered (above 80%, considered relatively sufficient).

Conclusion and Outlook. This poster presents our ongoing research of statistical ontology, which is easy to understand and covers most of statistical analytics in industrial applications. In the future we will improve on the design practices and try to reuse classes and properties (e.g., from STATO [13]) for better interoperability, some reasoning properties and mechanisms of the ontology, and further improve the purpose coverage.

Acknowledgements. The work was partially supported by the H2020 projects Dome 4.0 (Grant Agreement No. 953163), OntoCommons (No. 958371), and DataCloud (No. 101016835) and the SIRIUS Centre, Norwegian Research Council project number 237898.

References

[1] J. Davis, et al., Smart manufacturing, manufacturing intelligence and demand-dynamic performance, Computers & Chemical Engineering 47 (2012) 145–156.

[2] C. Naab, et al., Application of the unscented kalman filter in position estimation a case study on a robot for precise positioning, RobAutonSyst 147 (2022) 103904.
[3] E. C. Bryant, et al., Statistical analysis, 1966.
[4] Z. Zheng, B. Zhou, D. Zhou, et al., Executable knowledge graph for machine learning: A Bosch case for welding monitoring, in: ISWC, 2022.
[5] O. Celik, D. Zhou, et al., Specializing versatile skill libraries using local mixture of experts, in: CRL, PMLR, 2022, pp. 1423–1433.
[6] B. Zhou, Z. Zheng, D. Zhou, et al., The data value quest: A holistic semantic approach at bosch, ESWC, Springer (2022).
[7] D. Zhou, B. Zhou, et al., Ontology reshaping for knowledge graph construction: Applied on bosch welding case, in: ISWC, 2022.
[8] Z. Zheng, B. Zhou, et al., Query-based industrial analytics over knowledge graphs with ontology reshaping, ESWC, Springer (2022).
[9] D. Zhou, B. Zhou, Z. Zheng, et al., Enhancing knowledge graph generation with ontology reshaping–Bosch case, ESWC, Springer (2022).
[10] D. Zhou, et al., Schere: Schema reshaping for enhancing knowledge graph construction, in: CIKM, 2022.
[11] A. Salatino, et al., The computer science ontology: a large-scale taxonomy of research areas, in: ISWC, Springer, 2018, pp. 187–205.
[12] K. K. Breitman, et al., Ontology in computer science, Semantic Web: Concepts, Technologies and Applications (2007) 17–34.
[13] K. Kotis, A. Papasalouros, Statistics ontology, http://stato-ontology.org/, 2018.
[14] P. Rocca-Serra, et al., Experiment design driven fairification of omics data matrices, an exemplar, Scientific Data 6 (2019) 1–4.
[15] Z. Zheng, et al., Exekg: Executable knowledge graph system for user-friendly data analytics, in: CIKM, 2022.
[16] M. Yahya, et al., Towards generalized welding ontology in line with iso and knowledge graph construction, ESWC, Springer (2022).
[17] Z. Zheng, et al., Executable knowledge graph for transparent machine learning in welding monitoring at bosch, in: CIKM, 2022.
[18] B. Zhou, et al., Knowledge graph-based semantic system for visual analytics in automatic manufacturing, in: ISWC, 2022.
[19] B. Zhou, Z. Tan, et al., Towards a visualisation ontology for reusable visual analytics, in: IJCKG, 2022.
[20] N. F. Noy, et al., Ontology development 101: A guide to creating your first ontology, 2001.
[21] Z. Zheng, et al., Towards a visualisation ontology for data analysis in industrial applications, in: SemIIM@ESWC, 2022.
[22] B. Motik, et al., OWL 2 web ontology language profiles, 2012. URL: https://www.w3.org/TR/owl2-profiles/, accessed 5 July, 2022.