Ontology-based Design of Experiments on Big Data Solutions*

Maximilian Zocholl[1], Elena Camossi[1], Anne-Laure Jousselme[1], and Cyril Ray[2]

[1] NATO STO Centre for Maritime Research and Experimentation, La Spezia, Italy
{maximilian.zocholl,elena.camossi,anne-laure.jousselme}@cmre.nato.int
[2] Institut de Recherche de l'École Navale, Brest, France
{cyril.ray@ecole-navale.fr}

Abstract. In this paper the ontology-based approach is proposed to support the evaluation of big data systems. Firstly, the approach formalises a decomposition and recombination of the big data solution, allowing for the aggregation of component evaluation results at inter-component level. Secondly, existing work on Design of Experiments (DoE) is translated into an ontology for supporting the selection of experiments. It exploits domain and inter-domain specific restrictions on the factor combinations in order to select from the very large number of possible experiments a representative subset. Contrary to existing approaches, the proposed use of ontologies is not limited to the assertional description and exploitation of past experiments but offers richer terminological descriptions for the development of a DoE from scratch. As an application example, a DoE is developed for a maritime big data solution.

Keywords: Ontology · Big Data Solutions · Big Data Variations · Evaluation · Design of Experiments (DoE) · Situational Awareness

1 Introduction

The assessment of a big data system poses significant challenges in terms of the large number of data variations that must be considered by the experiments. To improve the evaluation efficiency, experiments must focus on a representative subset demonstrating the system ability to scale along the considered big data dimensions.

In many disciplines, Design of Experiments (DoE) is used to organise the evaluation of processes, systems, and products. DoE is a collective of principles, statistical approaches and models for planning and performing experiments as well as analysing their results [2, 6]. Typically, the experimental unit is modelled as a system with input and output variables. Some or all controllable input variables, the so called *factors*, are varied according to an experimental plan

* This work is supported by the Big Data Analytics for Time Critical Mobility Forecasting (datAcron) project, which has received funding from the European Unions Horizon 2020 research and innovation programme under Grant Agreement No. 687591.

that specifies the values of the variables, or *factor levels*, of each experiment. After feeding the experimental unit with a set of input variables, the output is observed. As the output depends on the system behavior and both controllable and possibly uncontrollable input variables, the goal of DoE is the quantification of the functional relation between the input and output of the system, e.g. by the analysis of variance or covariance. A large body of methods and knowledge exists, but barely formalized in a machine interpretable way.

Two main challenges arise during the assessment of a big data system with a DoE. Firstly, big data variations translate seamlessly into a large number of factors with a multiplicity of possible factor levels. Secondly, the different components of the big data system implement either deterministic or non-deterministic processes and yield different output types, e.g. continuous or multinomial. Thus, for a thorough assessment the system needs to be unfolded into its components. Again, this increases the number of necessary experiments but additionally and more importantly introduces the necessity of using completely different types of DoE. Choosing the wrong DoE results in a reduction of statistical efficiency or the lack of consistency of the results.

In this work, we apply ontology based DoE to the experimental evaluation of big data systems. The proposed formalisation encompasses the decomposition of the big data system, supporting the roll-up of the experiment results at component level to obtain the inter-component level evaluation. In addition, complementarily to the related work discussed in Section 2, in this paper we propose to expand the existing formalisations on DoE and leverage the domain knowledge to drive the selection of experiments. Similarly to [3], the knowledge is supposed to be captured in the T-Box of the Ontology, thus supporting DoE which starts without prior knowledge of the domain or instances in the A-Box.

2 Related Work

Do et al. provide an overview of empirical techniques for software testing identifying two approaches [1]: firstly, controlled experiments rely on the precise variation of given variables; complementarily, case studies follow possible scenarios of usage. Both approaches aim for the replicability of the performed experiments, the possibility to aggregate their results beyond the anticipated scope and by these means to validate the significance of their results in form of models. Precursors for enabling these benefits can be seen in the interpretability of the experimental results, e.g. by the documentation or standardisation, and in infrastructures allowing to share and connect artifacts [1, 5]. In the data mining and machine learning domain recent efforts lead to the W3C ML Schema Community Group with the goal to support the development of a data exchange standard for experimental data by unifying existing, more specific schemata [5]. More specifically, the group pools former efforts on Data Mining OPtimization (DMOP) [4], Expose, the OpenML related Ontology of Vanschoren [10], as well as the contributions of Soldatova et al. in the form of EXPO and OntoDM [9, 7], focussing on the process of scientific experiments. A very mature contribution is

WINGS, a semantic workflow system that eases the development of data mining workflows by its user friendly interface [3].

3 Ontology-based DoE and Evaluation

The goal of DoE is to choose an experimental plan that is statistically efficient whilst allowing for an aggregation of the experimental results that are consistent with the context of the experiments, as well as to accept or reject the research hypothesis. The choice of the DoE depends on multiple criteria, including the features of the experimental unit, the mathematical function to model the behavior of the experimental unit, the restrictions on the experimental space. The relation between these criteria and different DoE are modelled as ontology axioms, and the resulting ontology enables an ontology-based DoE. The following are a subset of the axioms modelled in OWL2:

$$DoEWithoutReplication \sqsubseteq DoE \sqcap \exists hasExpUnit.Deterministic. \quad (1)$$

$$DoEWithReplication \equiv DoE \sqcap \exists hasExpUnit.Deterministic. \quad (2)$$

$$NonDeterministic \equiv \exists hasComponent.NonDeterministic. \quad (3)$$

$$DoEWithBlocking \equiv DoE \sqcap \exists hasNuisanceFactor.Controllable. \quad (4)$$

$$DoEWithRandomization \equiv DoE \sqcap \exists hasNuisanceFactor.Uncontrollable. \quad (5)$$

For the evaluation of the ontology the maritime rule sets from [8] are used. The rule set is part of a big data solution currently developed in the execution of the datAcron project[3]. Spatial and critical rule sets are modelled as components of the composite rule sets.

Axiom (1) ensures only experimental units with deterministic behavior to be assigned to DoEs without replications. If an experimental unit with non-deterministic behavior is assigned to a DoE without replication, the ontology becomes inconsistent. All five rule sets in [8] are modeled as experimental units, namely “Vessel within Area”, “Vessel under Way”, “Aground”, “Trawling” and “Rendez-vous”. As all rule sets are deterministic, the DoE instances related to these different experimental units can be assigned correctly to the class DoEWithoutReplication. Assuming a non-deterministic behavior of an arbitrary component of a rule set, such as the user-driven detection of the same events in [8], axiom (2) and (3) induces an automatic classification of the respective DoE as instance of the class DoEWithReplication. In case of a rule set is assigned to the class Deterministic and a component of this rule is asserted as instance of the class Non-Deterministic, an inconsistency is created. As the Axioms (4) and (5) follow the same design pattern as (2), the available reasoning techniques presented in Table 1 allow for a similar support during the process of selecting a suitable DoE. All rule sets are deterministic and have no nuisance factors, neither controllable nor uncontrollable.

[3] www.datacron-project.eu

Table 1. Available Reasoning Techniques

DoE	Automatic assignment	Consistency check
DoEWithReplication	Yes	Yes
DoEWithoutReplication	No	Yes
DoEWithBlocking	Yes	Yes
DoEWithoutBlocking	No	Yes
DoEWithRandomization	Yes	Yes
DoEWithoutRandomization	No	Yes

4 Conclusion

As big data systems ingest a large number of variables with a large range of possible values, their evaluation requires a methodological choice of experiments. The presented approach uses the descriptions of components of an existing big data solution in order to reduce the design space of possible experiments according to well understood concepts of DoE. By excluding infeasible or unnecessary experiments, the number of experiments is reduced.

References

1. Do, H., Elbaum, S. and Rothermel, G.: Supporting Controlled Experimentation with Testing Techniques: An Infrastructure and its Potential Impact. Empirical Software Engineering **10**(4), 405–435 (2005).
2. Fisher, R. A.: The design of experiments. Oliver And Boyd, Edinburgh, London, (1937).
3. Gil, Y., Ratnakar, V., Kim, J., Gonzlez-Calero, P. A., Groth, P., Moody, J., Deelman, E.: WINGS: Intelligent Workflow-Based Design of Computational Experiments. Intelligent Systems, IEEE, (2011).
4. Keet, C.M., Ławrynowicz, A., d'Amato, C., Kalousis, A., Nguyen, P., Palma, R., Stevens, R. and Hilario, M.: The Data Mining OPtimization Ontology. Web Semantics: Science, Services and Agents on the World Wide Web **32**, 43–53, (2015).
5. ML Schema Core Specification, http://www.w3.org/2016/10/mls/. Last accessed: 8 Feb 2018.
6. Montgomery, D.C.: Design and Analysis of Experiments. John Wiley & Sons, (2017).
7. Panov, P., Soldatova, L., Džeroski, S.: Ontology of core data mining entities. Data Mining and Knowledge Discovery **28**(5), 1222–1265, (2014).
8. Pitsikalis, M., Kontopoulos, I., Artikis, A., Alevizos, E., Delaunay, P., Pouessel, J.-E., Dreo, R., Ray, C., Camossi, E., Jousselme, A.-L., Hadzagic, M.: Composite Event Patterns for Maritime Monitoring. In Proceedings of 10th Hellenic Conference on Artificial Intelligence SETN (2018).
9. Soldatova, L. N., King, R. D.: An ontology of scientific experiments. Journal of the Royal Society Interface **3**(11), 795–803, (2006).
10. Vanschoren, J., Blockeel, H., Pfahringer, B., Holmes, G.: Experiment databases. Machine Learning **87**(2), 127–158 (2012).