


Interactive Assistance for Scientific Workflow Modeling by Case-Based Reasoning

Christian Zeyen 

Business Information Systems II
University of Trier
54286 Trier, Germany
zeyen@uni-trier.de
<http://www.wi2.uni-trier.de>

Abstract Developing scientific workflows is a demanding task that can be supported in various ways. The range of approaches covers knowledge-intensive planning approaches to statistical approaches. Some of the key challenges are knowledge engineering and elicitation of the target problem to provide specific assistance. The presented research addresses these challenges with a case-based approach. The goal is to provide an interactive and self-improving assistant that collaborates with the developer not only to learn from newly created workflows but also to improve the underlying domain model.

Keywords: Case-Based Reasoning · Interactive Assistance · Scientific Workflows

1 Introduction

Scientific workflows [7] are designed for the computerized execution of data processing and analysis tasks. Scientific workflow management systems (cf. [6,3]) are powerful tools that provide graphical editors for composing workflows (also referred to as modeling) out of building blocks. However, developing scientific workflows remains a demanding task due to the large number of available components and the wide variety of possible compositions. Various assistance approaches exist that can be roughly divided into three groups: statistical (cf. [4]), planning-based (cf. [3]), and case-based approaches (cf. [2]). While semantic approaches typically require extensive knowledge engineering beforehand, less knowledge-intensive approaches such as statistical approaches need many available workflows to derive best-practice recommendations. Previous case-based approaches often come at high knowledge engineering costs and require a fully elaborated query by the user. Conversational approaches address the latter but cause additional effort for creating suitable dialogs. Moreover, previous work mainly focused on attribute-value representations of workflows and did not incorporate graph representations, which are typically used in workflow editors. Due to the complex nature of the domain, approaches are often limited to solve certain problems and require additional development work to keep up with continuously evolving

workflow systems. In general, it is challenging to assist modeling for the full range of composable workflows. For instance, it is common practice to use API calls or to perform arbitrary script execution within workflow execution. Many systems provide generic workflow components for this purpose, enabling to extend the functionality of built-in components. To provide an adequate assistance, it is essential to continuously refine the domain model and to involve workflow developers.

2 PhD Research Focus

Addressing the limitations of existing research, this PhD thesis research follows an interactive approach based on CBR to assist the modeling of scientific workflows. In a nutshell, based on the current workflow under development, a conversational retrieval is performed to find a suitable workflow serving as a template. Subsequently, applicable adaptations are suggested and performed in an interactive manner. Finally, a newly created workflow is stored as a new case including the semantic information obtained from the user interaction.

2.1 Research Questions

The following research questions will be addressed for the domain of scientific workflows:

1. How can a domain model be efficiently built based on available workflows and semi-structured meta data.
2. How can workflows be efficiently retrieved with a conversational approach?
3. How can interactive workflow adaptation be realized?
4. How can results be presented and explained to users?
5. How can knowledge required for conversational retrieval and adaptation be automatically derived from the case base and domain model?
6. How can user feedback be gathered and used to revise the knowledge model?
7. How can a conversational CBR approach be evaluated under real-world conditions?

2.2 Research Plan

The research is embedded in two projects. Research question 1 is investigated in the *eXplore!*¹ project, in which we build up the application domain and implement a basic case-based retrieval approach as an extension for the RapidMiner workflow system [6].

¹ *eXplore!* – *Computer-based Modeling, Analysis, and Exploration as a Basis for eScience in eHumanities* is a cooperation project with the Trier Center for Digital Humanities (TCDH) at the University of Trier.

Further work towards interactivity is done in the scope of the *EVER II*² project (cf. [1]). In the first place, a focus is put on the conversational retrieval of the workflow cases to investigate questions 2 and 4. For this purpose, we build up upon previous work [10,5]. In this context, important issues are time-efficient retrieval as well as question creation and sequence. Furthermore, previous evaluations [10,5] showed that an adequate presentation and explanation of results is essential (question 4). With respect to question 7, the idea is to deploy the research prototype for evaluating the assistance approach in public under real-world condition. By this means, usage data can be collected that may also lead to new insights into the process of workflow modeling.

Subsequent to conversational retrieval, the research activities focus on interactive adaptation (question 3). In this step, existing adaptation approaches will be integrated in the conversational framework. This step will also address questions 4 and 7.

An overall focus is put on reducing the initial effort for building such an assistance (question 1). Likewise the knowledge engineering effort at run-time for adapting the knowledge model to changing circumstances such as an evolving workflow system will be addressed. The research also addresses the knowledge acquisition bottleneck by deriving knowledge from the case base and the domain model (question 5) and by interactively acquiring knowledge during the workflow modeling process (question 6).

3 Current Progress

Previous work already addressed some of the research questions according to the research plan.

3.1 Application Domain

Question 1 was addressed during the implementation of the application domain. To assist workflow development under real-world conditions, the research is applied to the RapidMiner software [6] that allows for the visual programming of workflows for data and text mining tasks. RapidMiner is largely available under an open source license while also being distributed commercially, has an active user community, and is extensively expandable. In the *eXplore!* project, RapidMiner workflows were modeled for text processing and analysis tasks. In cooperation with digital humanists we investigated if workflow technology could be beneficial for humanities research following the model of eScience [8]. Within the project, we developed a prototypical modeling assistance as an extension for RapidMiner. The case-based approach supports the retrieval of RapidMiner workflows from a repository. A query for retrieval comprises the current workflow under development as well as keywords. The plugin also allows for extracting

² *EVER II – Extraction and Processing of Procedural Experiential Knowledge in Workflows – Quality, Interactivity, and Transferability* is a cooperation project with the Goethe University of Frankfurt

meta data about available workflow components and integrating the data into the knowledge model. For each such component, the model comprises various information such as textual descriptions, parameters, value ranges, default settings, input and output ports, and data types.

3.2 Conversational Retrieval of Workflows

Concerning questions 2, 4, and 5, our previous work [10] investigates an interactive retrieval approach that incrementally elicits the relevant features of the target problem. Thereby, we aim at reducing the effort and required expertise for the definition of queries. In contrast to other conversational approaches, cases are workflows that are represented as graphs. Questions are related to structural features and are automatically constructed based on extracted workflow fragments. Thereby, with respect to research question 5, the effort for defining suitable questions is omitted. An experimental evaluation with real users demonstrates that those features are meaningful subjects of questions and suitable to distinguish workflow cases from one another. The lessons learned from the evaluation are valuable for investigating question 7 in future work.

3.3 Query Model for Workflow Retrieval

In [5], addressing research question 2, we investigate expression elements to be used in a query language for scientific workflows. Based on a literature study, we present a query model consisting of workflow structure and meta description elements. The query model is evaluated with non-expert users in the RapidMiner workflow domain. It was observed in the experiments that the workflow structure is the most important query element followed by tags and keywords.

3.4 Automatic Adaptation of Workflows

In most recent work [9], we investigate automatic adaptation of scientific workflows as a first step towards answering question 3. With regard to our previous works on the adaptation of business workflows, we discuss differences between the workflow types and the resulting implications for transferring the adaptation approaches. We present two adaptation approaches namely substitutional adaptation by generalization and specialization of single workflow steps and structural adaptation with workflow streams that substitutes meaningful sub-components in workflows. The approaches learn the required adaptation knowledge from the case base, thus reducing the knowledge acquisition effort. An experimental evaluation demonstrates that both adaptation approaches can be used to significantly improve workflows towards a given query while mostly maintaining the executability and semantic correctness of the workflows. The work lays the foundation for interactive retrieval and adaptation of workflows that we consider to be key components of an interactive assistance.

References

1. Bergmann, R., Minor, M., Müller, G., Schumacher, P.: Project EVER: Extraction and Processing of Procedural Experience Knowledge in Workflows. In: Sánchez-Ruiz, A.A., Kofod-Petersen, A. (eds.) Proceedings of ICCBR 2017 Workshops. CEUR Workshop Proceedings, vol. 2028, pp. 137–146. CEUR-WS.org (2017)
2. Chinthaka, E., Ekanayake, J., Leake, D.B., Plale, B.: CBR Based Workflow Composition Assistant. In: 2009 IEEE Congress on Services, Part I, SERVICES I 2009, Los Angeles, CA, USA, July 6-10, 2009. pp. 352–355. IEEE Computer Society (2009)
3. Gil, Y., Ratnakar, V., Kim, J., González-Calero, P.A., Groth, P.T., Moody, J., Deelman, E.: Wings: Intelligent Workflow-Based Design of Computational Experiments. *IEEE Intelligent Systems* **26**(1), 62–72 (2011)
4. Jannach, D., Jugovac, M., Lerche, L.: Adaptive Recommendation-based Modeling Support for Data Analysis Workflows. In: Brdiczka, O., Chau, P., Carenini, G., Pan, S., Kristensson, P.O. (eds.) Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI 2015, Atlanta, GA, USA, March 29 - April 01, 2015. pp. 252–262. ACM (2015)
5. Malburg, L., Münster, N., Zeyen, C., Bergmann, R.: Query Model and Similarity-Based Retrieval for Workflow Reuse in the Digital Humanities. In: Gemulla, R., et al. (eds.) Proceedings of the Conference Lernen, Wissen, Daten, Analysen, LWDA 2018, Mannheim, Germany, August 22-24, 2018. CEUR Workshop Proceedings, vol. 2191, pp. 251–262. CEUR-WS.org (2018)
6. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: YALE: Rapid prototyping for complex data mining tasks. In: Proceedings of the 12th ACM SIGKDD Int. Conf. on Knowl. Discovery and Data Mining, 2006. pp. 935–940. ACM (2006)
7. Taylor, I.J., Gannon, D.B., Shields, M. (eds.): *Workflows for e-Science: Scientific Workflows for Grids*. Springer (2007)
8. Zeyen, C., Bergmann, R.: Towards an Interactive Workflow Modeling Assistance by Means of Case-Based Reasoning. In: Gemulla, R., et al. (eds.) Proceedings of the Conference Lernen, Wissen, Daten, Analysen, LWDA 2018. CEUR Workshop Proceedings, vol. 2191, pp. 355–360. CEUR-WS.org (2018)
9. Zeyen, C., Malburg, L., Bergmann, R.: Adaptation of Scientific Workflows by Means of Process-Oriented Case-Based Reasoning. In: Bach, K., Marling, C. (eds.) *Case-Based Reasoning Research and Development - 27th Int. Conf., ICCBR 2019*, Otzenhausen, Germany, September 8-12, 2019, Proceedings. LNCS, Springer (2019), Accepted for publication.
10. Zeyen, C., Müller, G., Bergmann, R.: Conversational Process-Oriented Case-Based Reasoning. In: Aha, D.W., Lieber, J. (eds.) *Case-Based Reasoning Research and Development - 25th Int. Conf., ICCBR 2017*, Trondheim, Norway, June 26-28, 2017, Proceedings. LNCS, vol. 10339, pp. 403–419. Springer (2017)