Towards an Interactive Workflow Modeling Assistance by Means of Case-Based Reasoning

Christian Zeyen and Ralph Bergmann

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

Abstract. Supporting non-experts in modeling workflows is an important yet challenging task to make workflow technology more accessible. While scientific workflows are broadly used in e-Science and various approaches to support modeling have been proposed, Digital Humanities (DH) is a rather new application domain that is largely unexplored, precisely in this respect. This paper presents our current work on developing an interactive modeling assistance for scientific workflows. We argue that interactivity is an important aspect when it comes to supporting the user in the iterative process of workflow modeling. We consider this process as a problem-solving activity and propose an assistance by case-based reasoning through the experience-based reuse of previously created workflows. We primarily target non-expert users and focus on text and data mining workflows, which we assume are in particular useful in the DH.

Keywords: Case-Based Reasoning · Workflow Reuse · Scientific Workflows

1 Introduction

Scientific workflows[15] can be regarded as an established means to perform experiments and data analysis in e-Science. Generally speaking, workflows are beneficial for the automation, modularization, reproducibility, reuse, documentation, and others. The potentials of scientific workflows have also been recognized in the context of the Digital Humanities (DH) although their broad usage is still in its infancy [12,13]. Particularly, modularization and reuse are topics of interest since various scientific tools[1] have been emerged from DH projects but reusing them for new research questions usually requires non-trivial and time-consuming adjustments or new combinations of tools. Scientific workflows particularly capture expert knowledge of how to solve a concrete problem in terms of required data, suitable processing steps and their composition, and parameter settings to name just a few. Hence, workflows can be valuable for non-expert users. By reusing workflows that have proven useful or repurposing them by experimenting with different combinations users may be able to perform non-trivial data analysis tasks

[1] This can be seen in the advent of tool and method collections such as TAPoR (http://tapor.ca) and Methodica (http://methodi.ca)

more efficiently [8]. Scientific Workflow Management Systems (SciWFM) typically enable the user to construct analytics processes at an more abstract level. For example, the RapidMiner (formerly known as YALE) [14] workflow editor supports the visual programming of workflows for data and text mining tasks. However, modeling new workflows can be a demanding and time-consuming task for novice users, in particular for complex data analysis that involve large amounts of data and require complex combinations of processing steps [6].

Consequently, supporting the development of scientific workflows has been considered an important topic of research and various approaches have been presented to this effect. Regarding RapidMiner, a recent feature [11] automatically creates data mining workflows for a user-defined data set and pre-defined analysis tasks. Another approach by Jannach et al. [9] supports the user with context-sensitive recommendations of suitable processing steps and parameter settings for the current workflow under construction based on the analysis of a large workflow repository. In [10], Kietz et al. summarize their findings from extending RapidMiner with semantic technologies. Building an ontology of the available processing steps, parameters, constraints, etc. enabled them to implement a fully automatic composition of workflows using a planning approach, correctness-checking of workflows, and quick-fixes to support users. Another SciWFM named WINGS [7] does also follow a semantic approach and uses planning and semantic reasoning to automatically create workflows based on high-level user requests. Case-Based Reasoning (CBR) [1] has also been utilized to support the development of workflows as an experience-based activity. For instance, Chinthaka et al. [5] proposed a generic approach that supports the keyword-based and graph-based search for workflows based on workflows annotated by the user beforehand.

Previous approaches that automatically compose or search workflows typically require the user to specify their requirements and analysis goals in a query. This poses the difficult dilemma of query elicitation. A query interface with a higher degree of abstraction might be more suitable for inexperienced users but is also more restrictive and thus less appropriate for users with more specific requests. On the other side, expressive query interfaces might be more suitable for expert users but the formulation of such a query can be a significant burden for the inexperienced user as it requires comprehensive domain knowledge. As a consequence, interactive search interfaces are considered to be important since they allow an iterative query refinement [6]. Further desirable features are the presentation of discriminative properties of similar workflows and the consideration of user feedback. Conversational CBR (CCBR) [2] approaches particularly target the problem of a proactive query formulation by conducting a dialog with the user. During a conversation, the user can answer consecutive questions instead of deciding proactively which information to include in the query. However, to our knowledge, CCBR has not yet been applied in the context of a modeling assistance for scientific workflows.

In this paper, we present ongoing work on interactive approaches to scientific workflow reuse. We particularly focus on supporting novice users such as domain scientists, students, or practitioners. More precisely, building up upon our previous works on process-oriented CBR, we are working on new interactive CBR approaches to a workflow modeling assistance. We plan to implement them as an extension to RapidMiner

and evaluate them in the domain of text and data mining workflows with non-expert users. We particularly aim to support digital humanists and we hope that our work can contribute to the establishment of scientific workflows in the DH.

In the following, section 2 describes our research goals and related projects in more detail while section 3 presents the proposed initial approach to an interactive workflow modeling assistance. Section 4 concludes with a short summary and outlook.

2 Research Goals and Projects

Our research is embedded within two projects located at the University of Trier. In the first project named *eXplore!*[2] we investigate the characteristics of workflow composition in the field of Digital Humanities. From the literary-scientific perspective, the goal is to investigate influences on the creative processes of Klaus Mann, a famous German writer, by analyzing documents about his life and experiences written by himself. A focus is put on his diaries from the 1930's that comprise very detailed information about his personal and professional life. Due to the wealth of information contained in the texts, text and data mining workflows are created and applied to extract, combine, and analyze the available data. Our primary goal is to accompany the creation of such scientific workflows and to develop a workflow modeling assistance suitable to support non-experts. The project is testing the use of RapidMiner as a SciWMS to create, apply, and manage workflows for the data analysis. In the course of the workflow creation, we capture the experiences made and we populate a structured workflow repository as a basis for further research on providing modeling support. Using CBR, we aim at providing an interactive workflow modeling assistance that facilitates the creation of new workflows by reusing and repurposing past ones.

For this purpose, we focus on developing new methods for the interactive retrieval and adaptation of workflows by means of process-oriented CBR (POCBR) in the second project named EVER[3]. During the first funding period from 2011 to 2016, the project has been investigating whether workflow technology and POCBR can help to analyze and reuse procedural experiential knowledge in Internet communities such as cooking web sites [4]. We considered recipes as cooking workflows and developed methods for the similarity-based retrieval of workflows and the automatic adaptation of retrieved ones. Throughout the project, we continuously integrated the developed methods into the CBR component of the CAKE framework [3]. Currently, the project is in its second funding period in which we focus on transferring the developed approaches to a novel application domain, i.e., text and data mining workflows. With regard to the goal of developing a workflow modeling assistance, *interactivity* will be a focus of research. The current retrieval and adaptation methods are fully automatic for a given user query and do not further interact with the user. To avoid the specification of user queries we proposed a conversational approach for the interactive retrieval of cooking workflows

[2] *eXplore! – Computer-based Modeling, Analysis, and Exploration as a Basis for eScience in eHumanities* is a cooperation project (launched in 2016) with the Trier Center for Digital Humanities (TCDH) at the University of Trier.

[3] *EVER – Extraction and Processing of Procedural Experiential Knowledge in Workflows* is a cooperation project (launched in 2011) with the Goethe University in Frankfurt.

[16]. However, adaptation has not yet been integrated. Our previous works revealed that the automatic adaptation approaches are capable of increasing the overall query fulfillment but potentially decrease the workflow quality. Consequently, we aim at developing new methods for the interactive retrieval and adaptation in order to better suit the user's needs. In particular, we hope that our approach is able to make workflow modeling more accessible for novice users.

3 Interactive Workflow Modeling Assistance by CBR

Figure 1 depicts the overall architecture of our case-based modeling assistance. Traditionally, the user creates, manages, and executes workflows in a graphical interface provided by a workflow management system. In order to facilitate the workflow creation, we integrate a CBR system into the user interaction. The integration comprises the transformation of workflows into semantic graphs and the formalization of domain knowledge about workflows and meta data into an ontology in order to enable the similarity assessment between workflows. Based on the transformed case base, adaptation knowledge can be learned automatically [4]. The actual modeling assistance is realized by implementing the CBR system following the well-known R^4-cycle [1] in order to *retrieve*, *reuse*, *revise*, and *retain* workflows.

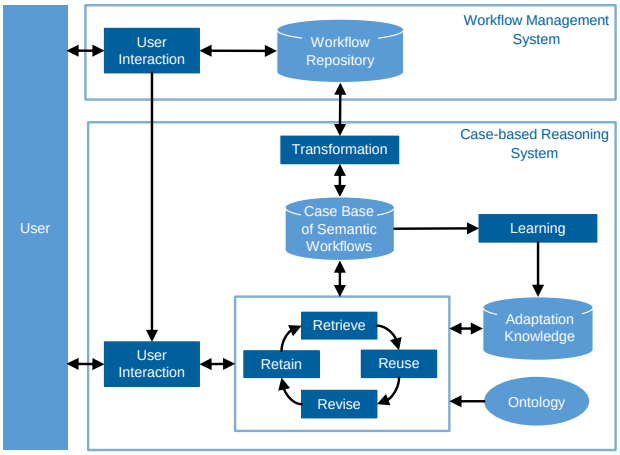

Fig. 1. Architecture of a Case-based Modeling Assistance

The interaction with the user plays an important role throughout the (iterative) workflow modeling process and the entire CBR cycle, respectively since we hope to obtain the following valuable information:

Retrieve At first, the user starts creating a workflow. Presumably, she does not yet know how the final workflow is constructed exactly and does only insert the input

and output data without specifying the entire data transformation. By accompanying the user's modeling process, information about the current *modeling context* can be gathered, e.g., the last workflow element the user inserted, and the retrieval can be invoked automatically whenever the workflow is updated. During the retrieve phase, the interaction component displays questions to the user and thus learns the user's *requirements and goals* successively. Based on the available information, an internal query is elaborated and best matching workflows from the repository are retrieved and ranked by similarity. The result list is displayed to the user.

Reuse In this phase, the user can apply her own modifications to the current workflow based on the workflows obtained from the retrieval. In addition, based on these workflows and the available adaptation knowledge, the CBR system computes applicable adaptations to the current workflow under construction and suggests them to the user. Such adaptations may consist of workflow fragments that can be (automatically) inserted into the current workflow. By this means, an autocompletion feature could be realized. Each adaptation must be selected by the user.

Revise This phase is closely linked to the reuse phase since each automatic adaptation may be undone or corrected by the user. Maintaining the quality of automatically learned adaptation knowledge is a challenging task. In this regard, the user can give valuable *feedback* to the system whether suggested or automatically performed adaptations have met the user's requirements.

Retain Retainment is essential for the learning of new *experience*. Whenever the user finalized a workflow and the workflow is tried and tested, she can inform the system about the new case and may also provide meta information that characterizes the specific application situation in which the workflow has proven useful. In addition, the CBR system can try to learn from manually applied adaptations that were performed by the user.

4 Summary and Outlook

In this paper we have presented ongoing work on the development of an interactive workflow modeling assistance by case-based reasoning. The target is to provide an assistance for non-experts that do not have a broad experience in creating workflows for text and data analysis. The question under investigation is whether interactive case-based reasoning can support inexperienced users (such as domain scientists or students) to familiarize themselves with the workflow programming paradigm and to create appropriate workflows in order to complete their desired analysis tasks. The Digital Humanities are a new and so far under-explored application domain with respect to this approach. In future work, we will focus on the interactive nature of problem-solving and investigate new methods for interactive retrieval and adaptation of already available workflows to provide users with useful suggestions for the current workflow under construction.

Acknowledgments. This work is partly funded by the German Federal Ministry of Education and Research (BMBF, No. 01UG1606) and the German Research Foundation (DFG, No. BE 1373/3-3).

References

1. Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications **7**(1), 39–59 (1994)
2. Aha, D.W., Breslow, L., Muñoz-Avila, H.: Conversational Case-Based Reasoning. Appl. Intell. **14**(1), 9–32 (2001)
3. Bergmann, R., Gessinger, S., Görg, S., Müller, G.: The collaborative agile knowledge engine CAKE. In: Goggins, S.P., Jahnke, I., McDonald, D.W., Bjørn, P. (eds.) Proceedings of the 18th International Conference on Supporting Group Work, 2014. pp. 281–284. ACM (2014)
4. 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)
5. 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. pp. 352–355. IEEE Computer Society (2009)
6. Cohen-Boulakia, S., Leser, U.: Search, Adapt, and Reuse: The Future of Scientific Workflows. SIGMOD Record **40**(2), 6–16 (2011)
7. 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)
8. Hauder, M., Gil, Y., Sethi, R.J., Liu, Y., Jo, H.: Making data analysis expertise broadly accessible through workflows. In: Taylor, I.J., Montagnat, J. (eds.) WORKS'11, Proceedings of the 6th Workshop on Workflows in Support of Large-Scale Science. pp. 77–86. ACM (2011)
9. Jannach, D., Jugovac, M., Lerche, L.: Supporting the Design of Machine Learning Workflows with a Recommendation System. TiiS **6**(1), 8:1–8:35 (2016)
10. Kietz, J.U., Serban, F., Fischer, S., Bernstein, A.: "Semantics Inside!" But Let's Not Tell the Data Miners: Intelligent Support for Data Mining. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Proceedings. LNCS, vol. 8465, pp. 706–720. Springer (2014)
11. Krishna Roy: RapidMiner looks to boost data scientists' productivity with Auto Model, https://rapidminer.com/resource/451-research-report-auto-model/
12. Kuhn, J., Reiter, N.: A Plea for a Method-Driven Agenda in the Digital Humanities. In: Book of Abstracts of DH 2015 (2015)
13. Kuras, C., Eckar, T.: Prozessmodellierung mittels BPMN in Forschungsinfrastrukturen der Digital Humanities. In: Eibl, M., Gaedke, M. (eds.) INFORMATIK 2017. pp. 1101–1112. Gesellschaft für Informatik, Bonn (2017)
14. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: YALE: Rapid prototyping for complex data mining tasks. In: Eliassi-Rad, T., Ungar, L.H., Craven, M., Gunopulos, D. (eds.) Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006. pp. 935–940. ACM (2006)
15. Taylor, I.J., Gannon, D.B., Shields, M. (eds.): Workflows for e-Science: Scientific Workflows for Grids. Springer London (2010)
16. 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 International Conference, ICCBR 2017, Proceedings. LNCS, vol. 10339, pp. 403–419. Springer (2017)