Short Report of the STSM

Curating Data Analysis Workflows for Better Workflow Discovery

Holder: Khalid Belhajjame, Université Paris-Dauphine, Paris, France
Host: Luciano Gerber, Manchester Metropolitan University, Manchester, UK.

Purpose of the STSM

Scientific Workflows can be large and complex, which hampers their understanding and ultimately their reuse. To tackle this problem, a handful of researchers have investigated the problem of workflow summarization. The obtained summary workflow should capture the essence of the experiments implemented by the corresponding executable workflows. For example, we showed in previous work how workflows can be summarized into smaller size workflows by using graph reduction rules. While useful, the solution we developed relies on semantic annotations that describes the tasks of the workflow, or more precisely the data manipulation they perform

In this STSM, we explored together with members of the Manchester Metropolitan University, namely Dr. Luciano Gerber and Dr. Raheel Nawaz, a different source of information that is readily available for summarizing workflows, namely textual description. Indeed, often the authors of scientific workflows provide texts describing the workflow before its publications. This is the case, for instance, for the workflows published on the popular myExperiment workflow repository [De Roure et al., 2009], which are accompanied with textual descriptions provided by the person who published the workflow. We posit in this work that textual descriptions can be leveraged to summarize scientific workflows.

Description of the work carried out during the STSM

During the STSM, we started by pondering the requirements that needs to be fullfiled by a workflow summary. After a careful examination of the state of the art, we identified the following requirements that need to be satisfied by workflow summaries.

- **Comprehensiveness**: The obtained summary needs to convey the necessary details for the user to understand the scientific experiment implemented by the original workflow
- **Minimality**: The number of tasks that compose the summary workflow needs to be as small as possible.
- **The summary is a workflow**: The graph-based nature of a workflow provides users with a means to quickly comprehend a workflow, in particular, when the workflow is of a small size.

Notice that the first two requirements are opposing in the sense that by seeking minimality once can compromise comprehensiveness and vice-versa.

We have, then, devised a solution that analyzes the textual description of the workflow with the objective of deducing a succinct workflow. In doing so, we used part of speech (POS) analysis of texts together with a technique that we elaborated to derive a graph out of the keyword extracted from the workflow description. We evaluated our method using 6 workflows that were published on the myExperiment website. The results of this analysis showed that while for certain workflows the description allowed the derivation of a sensible summary workflow, in other cases, the summary was completely unhelpful. This mixed results suggests the need for further investigation, which we started planning already (see next section).

Future collaboration with the host institution

The empirical result we have obtained compelled us to think of a new strategy. Indeed, we set out for a new sub-objective that we are currently looking at, which can be summarized as follows. “Can we assess the quality of a textual description and its informativeness”. We intend to look at this research question as part of our future research by using existing dependency graph techniques [DMM’08] that can be abducted from textual description. The expected finding will allow us to judge the quality of a description prior to ots use for abstracting (summarizing a workflow).

Foreseen publication

We intend to publish the results that will be come out of this research and the ongoing collaboration in the eScience conference.

References

[1] Pinar Alper, Khalid Belhajjame, Carole A. Goble, Pinar Karagoz:Small Is Beautiful: Summarizing Scientific Workflows Using Semantic Annotations. BigData Congress 2013: 318-325

[2] David De Roure, Carole A. Goble, Robert Stevens:The design and realisation of the myExperiment Virtual Research Environment for social sharing of workflows. Future Generation Comp. Syst. 25(5): 561-567 (2009)

[3] Marie-Catherine De Marneffe and Christopher D. Manning. "The Stanford typed dependencies representation." Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation. Association for Computational Linguistics, 2008.