Evaluating the results of Scientific Workflows based on provenance information and reproducibility analyzes

Anna Bánáti[1], Péter Kacsuk[2,3], Miklós Kozlovszky[1,2]

[1] Óbuda University, John von Neumann Faculty of Informatics, Biotech Lab
Bécsi str. 96/b., H-1034, Budapest, Hungary
[2] MTA SZTAKI, LPDS, Kende str. 13-17, H-1111, Budapest, Hungary
[3] University of Westminster, 115 New Cavendish Street, London W1W 6UW
emails: {banati.anna,kozlovszky.miklos}@nik.uni-obuda.hu

Keywords: scientific workflow, reproducibility analyses, provenance,

1. Introduction

Scientific workflows are often long running and compute intensive processes which are enacted on heterogeneous, parallel and distributed systems. The continuously changing nature of the environment make it hard or even prevent to reproduce the results or to share it in a scientist's community. On one hand provenance data have to be captured about the dataflow, the ancestry of results and the environment of the execution, on the other hand description data have to be collected from the scientist about the essential details, the types and samples of input/output data, and the operation of the experiment [1, 2]. All the information which is necessary to reproduce a workflow, we called descriptors. A workflow can be reproducible if all the descriptor's values are known and stored. However there can exist such descriptors which cannot be stored (for example too big input), can become unavailable in later time (for example volatile third party resources), vary in time (for example input originated from database which continuously changes) or are based on random generated values. In this case the full reproducibility is very challenges task or even impossible.

Currently the reproducibility of scientific workflows is a burning problem which the scientists and the system developers have to face with. Many researchers and developers deal with the implementation of tools or frameworks which facilitates reproducibility of the workflow (ReproZip, Vistrail, Research Object).

With our work we intend to support the scientist to create the reproducible workflow and we look for the answers for the questions like which part of the workflow will be reproducible, how much the likelihood of the reproducibility of the workflow or does the result of the workflow can be predictable or evaluable based on provenance information of previous executions. To achieve this goal we analyzed [3] the criteria of the reproducibility than we collected and categorized all the necessary information [4] and finally in this paper using certain existing ontologies [5] we give a statistical method to predict or evaluate the result. Based on this investigation we defined the descriptors of a workflow, the decay parameter of a descriptor and based on this definitions we set up a mathematical model of the reproducibility analyses to formalize the problem and determine the solution.

The ultimate goal of our work is to determine the procedure which can evaluate the results of a scientific workflow based on the different descriptors of the jobs and provenance

2. Description of a problem solution

Our approach is to investigate and analyze the availability and variability of all the collected information called descriptors which is necessary to reproduce the scientific workflows and we assign to each descriptors a so called decay parameter. The decay parameter describes the likeliness of the availability or the variability of a given descriptor's value. Accordingly, the value of decay parameter can be three different functions:

$$Decay(v) = \begin{cases} 0 & \text{if the descriptor's value is not changing in time;} \\ F(v), & \text{if the availablity of the given value varies according to the} \\ & \quad \text{probability distribution function } F(v) \\ vary(\Delta t, v), & \text{if the varying of the value is known} \end{cases} \quad (1)$$

If a workflow have successfully executed many time and we have collected provenance data about these executions we can create a sample set and we can analyze the changing descriptors consequently in certain cases we can evaluate it with a function. Sometimes this fluctuation of the descriptor value may be known at the beginning and the evaluation is unnecessary. In addition, analyzing the sample set not only the descriptor's values can be evaluated but even the result of the workflow.

3. Results

With help of statistical tools and a sample set created from provenance data we have gave a method to evaluate the result of the workflow when one of the descriptor's values are changing in time and the others are constant. Analyzing the sample set the fluctuation of the descriptors can be determined or estimated and then – applying the given vary function or estimation – the result of the reproduced workflow also can be determined or evaluated based on linear on non-linear regression depending on the sample set.

4. Conclusions and future work

We have analyzed the requirements of the reproducibility and investigated the varying of the descriptors of the scientific workflows in order to be able to predict the result or give a feedback to the scientist about the reproducibility ratio of his workflow. We have set up a mathematical model to formalize the problem and based on a sample set originated from provenance information about the previous execution of the scientific workflow we worked out a procedure to evaluate the results of the workflow re-executed in a later time. In our future work we would like to look for other tools and determine other procedures to give a more general solution for the problem when many descriptors' values can change simultaneously.

References

1. J. Zhao, J. M. Gomez-Perez, K. Belhajjame, G. Klyne, E. Garcia-Cuesta, A. Garrido, K. Hettne, M. Roos, D. De Roure, C. Goble: Why workflows break—Understanding and combating decay in Taverna workflows, in E-Science (e-Science), 2012 IEEE 8th International Conference on, 2012,
2. K. M. Hettne, K. Wolstencroft, K. Belhajjame, C. A. Goble, E. Mina, H. Dharuri, D. De Roure, L. Verdes-Montenegro, J. Garrido, és M. Roos, Best Practices for Workflow Design: How to Prevent Workflow Decay, in SWAT4LS, 2012
3. A. Banati, P. Kacsuk, M. Kozlovszky, M. Four level provenance support to achieve portable reproducibility of scientific workflows. In Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on IEEE
4. A. Banati, P. Kacsuk, M. Kozlovszky, "Minimal Sufficient Information about the Scientific Workflows to Create Reproducible Experiment," in *IEEE 19th International Conference on* Intelligent Engineering Systems, INES, 2015,
5. K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma, E. Mina, O. Corcho, J. M. Gómez-Pérez, S. Bechhofer G. Klyne C. Goble, Using a suite of ontologies for preserving workflow-centric research objects. Web Semantics: Science, Services and Agents on the World Wide Web. 2015