Future Generation Computer Systems 56 (2016) 584–594

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Effective and efficient similarity search in scientific workflow repositories

Johannes Starlinger [a,*], Sarah Cohen-Boulakia [b], Sanjeev Khanna [c], Susan B. Davidson [c], Ulf Leser [a]

[a] Humboldt-Universität zu Berlin, Institut für Informatik, Unter den Linden 6, 10099 Berlin, Germany

[b] Université Paris-Sud, Laboratoire de Recherche en Informatique, CNRS UMR 8623, INRIA, LIRMM, France

[c] University of Pennsylvania, Department of Computer and Information Science, 3330 Walnut Street, Philadelphia, PA 19104-6389, USA

HIGHLIGHTS

- We present a complete system for efficient scientific workflow similarity search.
- Workflow indexing integrates with repositories' existing search—no graphs needed.
- Layer Decomposition reranking of candidate workflows ensures high result quality.
- We evaluate on a large corpus of workflows with similarity ratings by human experts.
- Our system greatly improves previous results in speed and retains superior quality.

ARTICLE INFO

Article history:
Received 2 March 2015
Received in revised form
21 June 2015
Accepted 28 June 2015
Available online 3 July 2015

Keywords:
Scientific workflows
Similarity search

ABSTRACT

Scientific workflows have become a valuable tool for large-scale data processing and analysis. This has led to the creation of specialized online repositories to facilitate workflow sharing and reuse. Over time, these repositories have grown to sizes that call for advanced methods to support workflow discovery, in particular for similarity search. Effective similarity search requires both high quality algorithms for the comparison of scientific workflows and efficient strategies for indexing, searching, and ranking of search results. Yet, the graph structure of scientific workflows poses severe challenges to each of these steps. Here, we present a complete system for effective and efficient similarity search in scientific workflow repositories, based on the Layer Decomposition approach to scientific workflow comparison. Layer Decomposition specifically accounts for the directed dataflow underlying scientific workflows and, compared to other state-of-the-art methods, delivers best results for similarity search at comparably low runtimes. Stacking Layer Decomposition with even faster, structure-agnostic approaches allows us to use proven, off-the-shelf tools for workflow indexing to further reduce runtimes and scale similarity search to sizes of current repositories.

1. Introduction

Scientific workflow systems have become an established tool for creating and running reproducible in-silico experiments. With their increasing popularity, online repositories of scientific workflows have emerged as a means of facilitating sharing, reuse, and repurposing. Examples of such repositories include CrowdLabs [1], SHIWA [2], and the Galaxy repository [3]. Probably the largest workflow collection is myExperiment [4], which currently contains more than 2500 workflows from various disciplines, including bioinformatics, astrophysics, and earth sciences. To make the best use of these repositories, users need support to find workflows that match their specific needs [5]. However, currently these repositories only support keyword queries which are matched against textual descriptions, tags, and titles given to the workflows upon upload [2,3,1,4]. Obviously, the quality of such a search critically depends on the quality of the annotations associated with workflows.

Another source of information that can be exploited for search is the definition of a workflow itself [6]: scientific workflows

* Corresponding author.
E-mail addresses: starling@informatik.hu-berlin.de (J. Starlinger), cohen@lri.fr (S. Cohen-Boulakia), sanjeev@cis.penn.edu (S. Khanna), susan@cis.penn.edu (S.B. Davidson), leser@informatik.hu-berlin.de (U. Leser).

http://dx.doi.org/10.1016/j.future.2015.06.012
0167-739X/

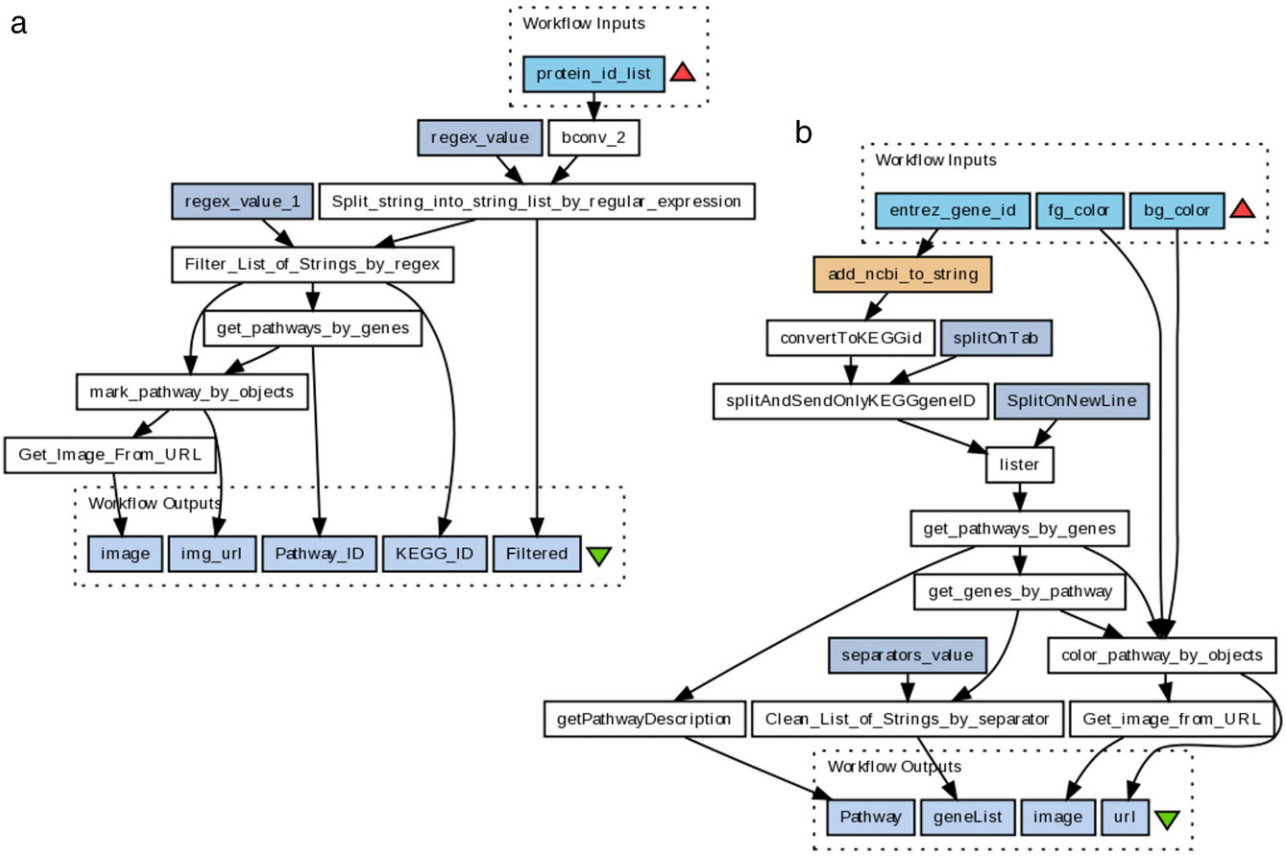

Fig. 1. Sample scientific workflows from the myExperiment repository: (a) ID: 1189, Title: *KEGG pathway analysis*, (b) ID: 2805, Title: *Get Pathway-Genes by Entrez gene id*.

typically resemble directed acyclic graphs (DAGs) consisting of global input and output ports, data processing modules, and datalinks which define the flow of data from one module to the next. Each module has a set of attributes associated with it, such as a descriptive label, the type of operation to be carried out, or, for instance, the uri of a web-service to be invoked. Two sample workflows from myExperiment are shown in Fig. 1. This structure or topology of the workflow, together with the attributes defined on the workflow's modules, is used by structure-based methods of workflow comparison. An obvious advantage of structure-based approaches to workflow similarity search is that they do not require any additional information to be provided by the workflow designer apart from the workflow itself. Structure-based approaches are typically used in a second search phase: first, users identify workflows which roughly match their needs using keyword search. In the second phase, users select one candidate workflow and let the system retrieve functionally similar workflows, i.e., the system performs a workflow similarity search.

Several studies have investigated different techniques for assessing workflow similarity using this attribute-enriched structure [7,8,6,9–12], but initial results indicate that they perform no better, and sometimes even worse, than annotation-based methods in terms of retrieval quality [13,8,11]. However, these comparisons were performed on very small and well-documented workflow sets, and thus the results should not be extrapolated to the large, but shallowly annotated repositories that exist today. To verify this hypothesis, in prior work we performed a large-scale comparative evaluation of workflow similarity search algorithms [14]. Our results indicated that (a) structure-based methods are indispensable for some current repositories which lack rich annotations, (b) structure-based methods, once properly configured, outperform annotation-based methods even when such rich annotations are available, and (c) any such standalone approach is further beaten by ensembles of annotation-based and structure-based methods. We also discovered that both the amount of configuration required and runtime considerations were drawbacks to such methods: fast workflow comparison using annotations on the workflows' modules provides best results only when ubiquitous, functionally unspecific modules are removed from the workflows in a preprocessing step. The configuration of which modules are to be removed is specific to a given dataset, and is non-trivial. Methods based on workflow substructures, on the other hand, provide rather stable results across different configurations, but have prohibitive runtimes.

Based on these findings we presented a novel technique for measuring workflow similarity that accounts for the directed dataflow underlying scientific workflows [15]. The central idea is the derivation of a *Layer Decomposition* for each workflow, which is a compact, ordered representation of its modules, suitable for effective and efficient workflow comparison. Comparatively evaluating this novel technique against previous approaches, we showed that the algorithm (a) delivers the best results in terms of retrieval quality when used stand-alone, (b) is essentially configuration free which makes it applicable to any workflow repository, regardless of how well its workflows are annotated, (c) is faster than other algorithms that account for the workflows' structure, and (d) can be stacked and combined with other measures to yield better retrieval at even higher speed.

Extending on these encouraging results we here investigate their transferability into a system for fast similarity search for scientific workflows at repository-scale. While runtime has been a concern in developing the *Layer Decomposition* approach, scaling its quality of scientific workflow comparison to large collections of workflows requires additional considerations as of how to best index workflows for fast retrieval. Especially our previous findings regarding the stackability of Layer Decomposition with other