SCIENCE CHINA
Information Sciences

• MOOP •

November 2016, Vol. 59 113101:1–113101:4
doi: 10.1007/s11432-015-0934-9

Similarity assessment for scientific workflow clustering and recommendation

Zhangbing ZHOU[1,2*], Zehui CHENG[1] & Yueqin ZHU[3]

[1]*School of Information Engineering, China University of Geosciences (Beijing), Beijing* 100083*, China;*
[2]*Computer Science Department, TELECOM SudParis, Evry* 91011*, France;*
[3]*Development Research Center of China Geological Survey, and Key Laboratory of Geological Information Technology, Ministry of Land and Resources, Beijing* 100037*, China*

Received May 18, 2016; accepted August 22, 2016; published online October 14, 2016

Abstract This article proposes to identify and recommend scientific workflows for reuse and repurposing. Specifically, a scientific workflow is represented as a layer hierarchy that specifies the hierarchical relations between this workflow, its sub-workflows, and activities. Semantic similarity is calculated between layer hierarchies of workflows. A graph-skeleton based clustering technique is adopted for grouping layer hierarchies into clusters. Barycenters in each cluster are identified, which serve as core workflows in this cluster, for facilitating the cluster identification and workflow ranking and recommendation with respect to the requirement of scientists.

Keywords scientific workflow, reuse and repurposing, layer hierarchy, graph-skeleton, recommendation

Citation Zhou Z B, Cheng Z H, Zhu Y Q. Similarity assessment for scientific workflow clustering and recommendation. Sci China Inf Sci, 2016, 59(11): 113101, doi: 10.1007/s11432-015-0934-9

1 Introduction

To facilitate the reuse and repurposing of data-intensive scientific processes, recurring data and computational resources are wrapped as web services nowadays [1], which are publicly accessible to the others through standard interfaces [2], and assembled as scientific workflows. Several online repositories have been emerged recently to allow the sharing of scientific workflows, including myExperiment wherein a large number of scientific workflows from various disciplines have been published. Along with an increasing number of scientific workflows published and shared on the Web, scientists can reuse or repurpose (partial of) current workflows, rather than developing from scratch, when the requirement may be (partially) achieved by (sub-)workflows which are evident as best practices in the past. Note that creating scientific workflows from scratch is a knowledge-intensive and laborious task. Hence, the reuse and repurposing of current scientific workflows is a promising strategy for constructing mission-oriented workflows.

Traditional techniques have studied the similarity assessment of scientific workflows [3] for facilitating their reuse and repurposing, and these techniques can be categorized into structure- or annotation-based strategies. Generally, the layer-hierarchy in scientific workflows has not been explored extensively. This means that the similarity assessment may not be accurate somehow. Considering the layer-hierarchy

* Corresponding author (email: zhangbing.zhou@gmail.com)

© Science China Press and Springer-Verlag Berlin Heidelberg 2016

for computing the similarity, and thus ranking and recommending appropriate scientific workflows, for facilitating the reuse and repurposing of current workflows, is a challenge to be explored further.

To address this challenge, we propose a novel technique for assessing the similarity, considering the layer-hierarchies of scientific workflows, besides their structures and text descriptions. Our major contributions are summarized as follows: (i) Semantic similarity for scientific workflows is computed leveraging their representation as layer-hierarchies, such that the hierarchical relations about a workflow, its sub-workflows and activities are considered, besides the textual and structural description of workflows. (ii) Scientific workflows are grouped into clusters using a graph-based clustering technique [4]. The barycenters identified in a cluster are regarded as the representatives of this cluster, for facilitating the cluster identification and workflow ranking and recommendation w.r.t. the requirement of scientists.

2 Preliminaries

A scientific workflow swf is a tuple (tl, dsc, SWF_{sub}, ACT, LNK), where tl is the title, and dsc is the text description, of swf. SWF_{sub} is a set of sub-workflows contained in swf. ACT is a set of activities contained in swf. LNK is a set of data links that connect sub-workflows in SWF_{sub} and activities in ACT = {(nm, dsc)}. It is worth noting that a sub-workflow can be an independent scientific workflow as well. In fact, a scientific workflow can be reformatted as a layer-hierarchy, leveraging sub-workflows and activities wherein. A layer-hierarchy lhr_{swf} of a scientific workflow swf is a tuple (tl, dsc, SWF_{sub}, ACT, LNK, LNK_{lh}), where LNK_{lh} is a set of links connecting sub-workflows and their specifications. We present the procedure for computing the semantic similarity of two activities $\text{act}_1 = (\text{nm}_1, \text{dsc}_1, \text{LNK}_1)$ and $\text{act}_2 = (\text{nm}_2, \text{dsc}_2, \text{LNK}_2)$ as follows.

- The minimum cost and maximum flow algorithm and WordNet are adopted for computing the similarity between names of activities $\text{sim}_{\text{actNm}}(\text{act}_1.\text{nm}_1, \text{act}_2.\text{nm}_2)$, and an algorithm xsimilarity is used to compute the similarity between text descriptions of activities $\text{sim}_{\text{actDsc}}(\text{act}_1.\text{dsc}_1, \text{act}_2.\text{dsc}_2)$.
- The semantic similarity for two activities is calculated through

$$\text{sim}_{\text{act}}(\text{act}_1, \text{act}_2) = \varrho \times \text{sim}_{\text{actNm}}(\text{act}_1.\text{nm}_1, \text{act}_2.\text{nm}_2) + (1 - \varrho) \times \text{sim}_{\text{actDsc}}(\text{act}_1.\text{dsc}_1, \text{act}_2.\text{dsc}_2), \quad (1)$$

where the factor $\varrho \in [0, 1]$ reflects the relative importance of $\text{sim}_{\text{actNm}}$ w.r.t $\text{sim}_{\text{actDsc}}$. Generally, $\text{sim}_{\text{act}}(\text{act}_1, \text{act}_2)$ returns a value between 0 and 1. The bigger the sim_{act} is, the more similar the activities act_1 and act_2 are.

We have examined the activities in scientific workflows located within Taverna 1/2 in the myExperimemt repository. A name may be an abbreviation which is not a valid word (like "FASTA"), and thus, cannot be recognized by WordNet. When these cases occur, ϱ is set to 0 for not considering the name.

3 Workflow network model construction

Leveraging the semantic similarity between scientific workflows, a scientific workflow network model WfN is constructed. A scientific workflow model is specified as a tuple (SWF, LNK, WGT), where SWF is a set of scientific workflows, LNK is a set of links that connect scientific workflows contained in SWF, and WGT is a set of weights defined upon LNK, which specifies the similarity between scientific workflows.

An example of workflow network model is shown in Figure 1, where the nodes refer to scientific workflows, and the weights upon the edges represent the similarity between activities.

Generally, the similarity between two scientific workflows is computed leveraging their layer hierarchies (i.e., $\text{lhr}_{\text{swf1}} = (\text{tl}_1, \text{dsc}_1, \text{SWF}_{\text{sub1}}, \text{ACT}_1, \text{LNK}_1, \text{LNK}_{\text{lh1}})$ and $\text{lhr}_{\text{swf2}} = (\text{tl}_2, \text{dsc}_2, \text{SWF}_{\text{sub2}}, \text{ACT}_2, \text{LNK}_2, \text{LNK}_{\text{lh2}})$). The computational steps are presented as follows.

- Step 1. Similarity computation for pairwise activities in $\text{lhr}_{\text{swf1}}.\text{ACT}_1$ and $\text{lhr}_{\text{swf2}}.\text{ACT}_2$, through the mechanism specified by Formula (1). A vector $\text{ACT}_{\text{sim}} = \{(\text{act}_1, \text{act}_2, \text{sim}_{\text{act}}(\text{act}_1, \text{act}_2))\}$ is adopted for recording the value of similarity for two activities $\text{act}_1 \in \text{lhr}_{\text{swf1}}.\text{ACT}_1$ and $\text{act}_2 \in \text{lhr}_{\text{swf2}}.\text{ACT}_2$.

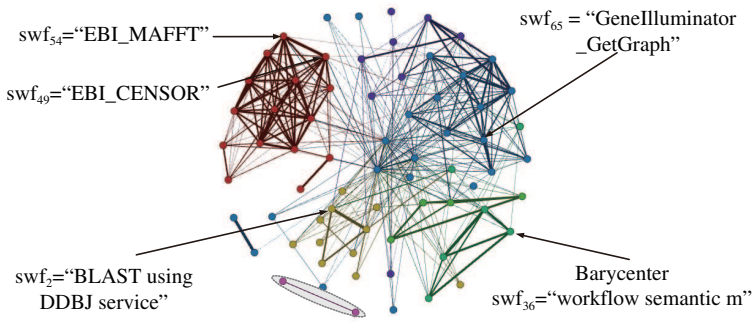

Figure 1 (Color online) A workflow network model with 69 nodes and 364 edges. There are 6 clusters generated whose nodes and edges are marked using different colors. The node swf_{36} is an example of the barycenter.

- Step 2. Given a layer hierarchy $\text{lhr}_{\text{swf1}} = (\text{tl}_1, \text{dsc}_1, \text{SWF}_{\text{sub1}}, \text{ACT}_1, \text{LNK}_1, \text{LNK}_{\text{lh1}})$ for a scientific workflow, a pre-order traversal node list (denoted $\text{NDS}_{\text{prodr}}$) and a post-order traversal node list (denoted $\text{NDS}_{\text{pstodr}}$) are generated through a depth-first search algorithm.
- Step 3. Similarity is computed for the pre-order traversal node lists (i.e., $\text{NDS}_{\text{prodr}}$) for the layer hierarchies of scientific workflows. Note that the strategy for processing the post-order traversal node lists (i.e., $\text{NDS}_{\text{pstodr}}$) is the same as that of pre-order traversal node lists, and we do not detail it. The similarity for $\text{NDS}^1_{\text{prodr}}$ and $\text{NDS}^2_{\text{prodr}}$ is computed as follows:

$$\text{sim}_{\text{prodr}}(\text{NDS}^1_{\text{prodr}}, \text{NDS}^2_{\text{prodr}}) = \frac{1}{2} \times (\text{sim}_{\text{lst}}(\text{NDS}^1_{\text{prodr}}, \text{NDS}^2_{\text{prodr}}) + \text{sim}_{\text{lst}}(\text{NDS}^2_{\text{prodr}}, \text{NDS}^1_{\text{prodr}})). \tag{2}$$

$\text{sim}_{\text{prodr}}()$ returns a value between 0 and 1, where 0 means totally different, while 1 means equivalent.

- Step 4. Finally, the similarity for the layer hierarchies of scientific workflows is computed as follows:

$$\text{sim}_{\text{lhr}}(\text{lhr}_{\text{swf1}}, \text{lhr}_{\text{swf2}}) = \min(\text{sim}_{\text{prodr}}(\text{NDS}^1_{\text{prodr}}, \text{NDS}^2_{\text{prodr}}), \text{sim}_{\text{pstodr}}(\text{NDS}^1_{\text{pstodr}}, \text{NDS}^2_{\text{pstodr}})), \tag{3}$$

where the function min() returns the minimum from a set of values. After computing the similarity between layer hierarchies, a workflow network model is constructed. An example is shown in Figure 1, where there are 69 scientific workflows corresponding to the nodes in this model.

4 Workflow clustering and recommendation

4.1 Graph-skeleton-based clustering

This section aims to group scientific workflows, which correspond to the nodes in a workflow network model (denoted WfN), into clusters. A graph-skeleton-based clustering (gSkeletonClu) algorithm is adopted [4], which is a density-based network method, for clustering nodes in WfN. When this algorithm is applied, hierarchial clusters, hubs, and outliers are determined. We use the notion CLS to denote the set of clusters generated. Intuitively, a cluster $\text{cls} = (V_{\text{cls}}, E_{\text{cls}}, \varpi_{\text{cls}}) \in \text{CLS}$ contains a set of nodes V_{cls}, and a set of edges connecting these nodes E_{cls}. The structural similarity computed by this algorithm for a hub hb and a node in cls. V_{cls} is retrieved and summed up. Thereafter, the cluster cls_{sel} with the largest average structural similarity is determined and assumed as the cluster that should contain hb. Consequently, hb is inserted into $\text{cls}_{\text{sel}}.V_{\text{cls}}$, and the edges with the structural similarity are handled accordingly. The assignment procedure of outliers is the same as those for the hubs.

Generally, workflows within a cluster are similar in functionality, and can be used for the reuse and repurposing when a similar requirement is presented by scientists.

4.2 Barycenter determination for clusters

With the clusters generated in the previous section for the grouping of scientific workflows, we propose to identify core workflows to represent a certain cluster, which correspond to the most representative

workflows in this cluster. Without loss of generality, the barycenters are identified for representing core workflows, where a barycenter of a weighted graph refers to a node in this graph, where the sum of the weights specified upon the edges connecting this node is among the largest.

Given a cluster cls = (V_{cls}, E_{cls}, ϖ_{cls}) and a workflow network model WfN = (SWF, LNK, WGT), the barycenters are determined as follows: Given a node $v \in V_{\text{cls}}$, we get a set of edges E^v_{cls} in E_{cls} connecting v, and compute the sum of weights (denoted wgt^{bc}_v) upon E^v_{cls} with respect to the weights specified in WfN.WGT. The nodes in V_{cls} are sorted in a descending order according to their values of wgt^{bc}_v. Therefore, the nodes, whose values of wgt^{bc}_v are within the top tp% (e.g., 33%), are chosen as the barycenters of cls. For instance, a barycenter is marked as swf_{36} = "workflow semantic m" in Figure 1.

4.3 Scientific workflow ranking and recommendation

Leveraging the barycenters identified in clusters aforementioned, we propose to identify the most appropriate cluster, where workflows within the cluster are examined, ranked, and recommended to scientists. The procedure of this cluster identification and workflow recommendation task is presented as follows, where lhr_{usr} represents a layer hierarchy with respect to the scientist's requirement.

- Step 1. Given a cluster cls where one of whose barycenters is denoted as $\text{bc}_{\text{cls}} \in$ cls.BC, the similarity (denoted $\text{sim}_{\text{bc}_{\text{cls}}}$) for bc_{cls} and lhr_{usr} is computed through Formula (3).
- Step 2. A cluster cls is assumed as the cluster cls_{sel} to be selected, where layer hierarchies contained are the most similar to lhr_{usr}. This means that the average similarity for barycenters in cls is the largest.
- Step 3. After the determination of the candidate cluster cls_{sel}, candidate layer hierarchies are determined accordingly, where the similarity for v_{cls} and lhr_{usr} is computed through Formula (3).
- Step 4. Layer hierarchies, whose similarity with lhr_{usr} is among top k% in values, are assumed the most beneficial for the development of lhr_{usr}, and thus are recommended for reuse and repurposing.

5 Conclusion

This article proposes a novel technique to promote the reuse and repurposing of scientific workflows, where a layer hierarchy is adopted to represent the hierarchical relations between a workflow, its sub-workflows, and activities. When a layer hierarchy is presented to reflect the requirement of scientists, a cluster is identified, such that the workflows in this cluster are ranked and recommended, for facilitating the development of a novel scientific workflow with respect to this requirement of scientists.

Acknowledgements This work was supported partially by National Natural Science Foundation of China (Grant Nos. 61379126, 61662021).

Conflict of interest The authors declare that they have no conflict of interest.

Supporting information The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Liu X Z, Huang G, Zhao Q, et al. Imashup: a mashup-based framework for service composition. Sci China Inf Sci, 2014, 57: 012101
2 Ning H S, Liu H. Cyber-physical-social-thinking space based science and technology framework for the internet of things. Sci China Inf Sci, 2015, 58: 031102
3 Starlinger J, Brancotte B, Cohen-Boulakia S, et al. Similarity search for scientific workflows. Proc VLDB Endowment, 2014, 7: 1143–1154
4 Huang J, Sun H, Song Q, et al. Revealing density-based clustering structure from the core-connected tree of a network. IEEE Trans Knowl Data Eng, 2013, 25: 1876–1889