Information Sciences 314 (2015) 1–13

Contents lists available at ScienceDirect

Information Sciences

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

# Cost and accuracy aware scientific workflow retrieval based on distance measure

Yinglong Ma [a,b,*], Moyi Shi [a], Jun Wei [b]

[a] School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, PR China
[b] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, PR China

ARTICLE INFO

*Article history:*
Received 24 May 2013
Received in revised form 16 March 2015
Accepted 20 March 2015
Available online 30 March 2015

*Keywords:*
Scientific workflow
Workflow retrieval
Distance measure
Matrix representation

ABSTRACT

Scientific workflows have been applied in many scientific areas with the large amount of complex data computation tasks such as life science, astronomy and earth science, etc. However, most existing approaches for scientific workflow retrieval neglect some constraints of quality of services (QoS) that users are really concerned about, and fail to allow users to express and retrieve scientific workflows with arbitrary constraints based on graph structures of workflows. In this paper, we propose a novel approach for scientific workflow retrieval with cost constraints. We present a graph representation model called Cost Constrained Graph (CCG) for representing scientific workflows with cost constraints. A distance measure is defined for accurate workflow retrieval. The CCGs representing candidate workflows can be ranked by comparing the similarity among them. We also theoretically prove that this measure satisfies all the four properties of distance. Furthermore, we develop a prototype system for editing, assignment of weights, and automatic similarity computation of workflows. At last, the related experiments are made to demonstrate the usefulness and efficiency of workflow retrieval based on our approach.



## 1. Introduction

We have witnessed continued growth in scientific workflow applications over the last decade. Scientific workflows have been applied in many scientific areas with the large amount of complex data computation tasks [35], such as bioinformatics, astronomy, ecology, earth science, etc. Considerable effort has been made for the development of scientific workflow management systems. A variety of scientific workflow systems have been developed such as Kepler [31], Taverna [37], Pegasus [21], Triana [16], Vistrail [38], and Pipeline Pilot [55], accelerating the pace of scientific progress in these scientific areas.

Scientific workflows are used to represent and manage complex distributed computations and data manipulation steps [46]. The users of scientific workflows involve not only computer scientists of developing them, but also the domain scientists of using them for science. The domain scientists are the "true users" in using scientific workflow systems [42]. They are indeed concerned about how to apply advanced methods on new data for discovering new facts in their respective science. Unfortunately, the domain scientists in real life usually do not know how to develop a scientific workflow for applying these advanced methods, and even do not know how to program. However, we argue that scientific workflows are developed with the aim of scientific experimentation and can be "repeatedly" executed with different data or different parameters [23]. The

* Corresponding author at: School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, PR China. Tel.: +86 10 61772643.
*E-mail address:* yinglongma@gmail.com (Y. Ma).

http://dx.doi.org/10.1016/j.ins.2015.03.055
0020-0255/

fact that new discoveries are achieved at an impressive speed in many scientific fields such as Life Sciences does not imply that each discovery would have a new method accompanied [17]. The methods used in scientific experiments are often highly similar, so the domain scientists for scientific data analysis often focus on producing large amounts of high-quality experimental data instead of deeply understanding these methods. Once having the data, the domain scientists would be highly interested in searching a repository of workflows for analyzing them by using the best possible methods. Essentially, scientific workflow retrieval has become a promising solution for domain scientists to find a suitable candidate workflow satisfying their requirements to a given problem instead of developing one themselves. Retrieving workflows can often be done by matching the expressed constraints between candidate workflows and the requirements of users in terms of their quality of services (QoS) such as cost, responding time and reliability, etc. Currently, some approaches for workflow retrieval were proposed such as Business Process Query Language (BPQL) [5] and BPMN-Q [2], which have demonstrated potential and initial success in business workflow retrieval.

However, some important open issues for scientific workflow retrieval remain to be resolved. First, the cost of services in workflows is neglected during the scientific workflow retrieval. A complex scientific workflow for scientific data analysis is composed of dozens of tasks/services which are often provided by external service providers. Users often need to pay for service access (although some of services are free to access, their QoS such as availability may not be guaranteed). In the case, users have to make a reservation with a service provider in advance to guarantee the service availability. Different service providers may provide different service costs and different levels of QoS. A service with the same functionality often has different QoS levels corresponding to different costs. Therefore, users can negotiate with service providers on service level agreements for required QoS. The price of a service can be determined by the processing speed, time and cost of the service. In general, service providers can charge higher prices for higher QoS. However, users may not always need to complete workflows earlier. They sometimes may prefer to use cheaper services with a lower QoS that is sufficient to meet their requirements. Unfortunately, there are seldom approaches and models of retrieving scientific workflows based on QoS requirements such as service cost, etc.

Second, most existing approaches fail to allow users to express arbitrary constraints based on graph structures of workflows. This will impede the customization and reuse of scientific workflows. For example, if users want to explicitly search whether a workflow contains loops and branches that satisfy their cost constraints, graph structures of workflows must be considered for such searches. However, on one hand, there is no such graph model for scientific workflows to represent their graph structures with the cost constraints. Approaches for retrieving workflows by comparing structures with cost constraints have not been found. Comparing graph structures allows us to match and rank candidate workflows by similarity computation with the cost constraints. On the other hand, the domain scientists are possibly not experts in any query language. There is a lack of tools to be able to express their retrieval requests as easily as possible and support a rich set of graph edit operations (e.g., adding/removing/replacing of a flow or a task), assignments of weights based on the cost constraints, and automatic similarity computation of scientific workflows.

In this paper, we propose a novel approach for scientific workflow retrieval with cost constraints. First, we present a graph representation model called Cost Constrained Graph (CCG) for representing scientific workflows with the cost constraints of QoS. Second, we define a distance measure for comparing the similarity among CCGs through which workflow retrieval and ranking can be made based on similarity computation. This measure is theoretically proved to satisfy all the four properties of distance. Third, we develop a tool for editing, assignment of weights, and automatic similarity computation of workflows. At last, the related experiments are made to demonstrate the usefulness and efficiency of workflow retrieval based on our approach. We argue that our approach can be also applied to model and retrieval workflows with other QoS requirements such as response time and speed.

This paper is organized as follows. Section 2 reviews the existing work. In Section 3, we give an overview of our CCG based approach. Section 4 introduces the basic notations related to CCG and cost constraints. Section 5 is to discuss the normalized matrix representations for two workflows to be compared, and introduces the distance measure. We also theoretically prove that it satisfies all properties of distance. Section 6 discusses how to compare workflows by some algorithms developed. Section 7 is the related experiments and evaluation. Section 8 is the conclusion and the future work.

## 2. Related work

Reusing scientific workflows has received more attention in recent years [24,43,10,40,41,39,29,32,49,53,54,20,48], and retrieval of processes provides a feasible solution for their reuse [25]. Users of scientific workflows would like to simply choose one of the matching workflows from workflow repository, provide its input data, and obtain the output results. A local scientific workflow management system can take the input data, download and execute the selected workflow, and obtain the output. The executed and traced workflow can be also uploaded to the repository for later reuse.

Some research was done in expressing what users of scientific workflows would like to retrieve. Recently, most existing approaches describe their data and intended analysis by using a set of keywords [37,30]. These input keywords are matched against the candidate workflows in the repository according to their documentation, metadata, data types, task names, etc. Keywords based workflow retrieval does not consider the structures of scientific works, and therefore cannot really satisfy users retrieval requirements. Some approaches with more expressivity were developed for workflow query, such as BPQL [5] and BPMN-Q [2]. However, users are often not the experts in process query languages, and the use of these languages is not easy to them. Moreover, whether or not these approaches for *workflow query* in traditional business processes management