Provenance-Based Searching and Ranking for Scientific Workflows

Víctor Cuevas-Vicenttín[1](✉), Bertram Ludäscher[1], and Paolo Missier[2]

[1] Department of Computer Science, University of California at Davis,
One Shields Avenue, Davis, CA 95616, USA
victorcuevasv@gmail.com, ludaesch@ucdavis.edu

[2] School of Computing Science, Newcastle University, Claremont Tower 9.08,
Newcastle upon Tyne NE17RU, UK
paolo.missier@ncl.ac.uk

Abstract. We present PBase, a scientific workflow provenance repository that supports declarative graph queries and keyword-based graph searching, complemented with ranking capabilities taking into consideration authority and quality of service criteria. Given the widespread use of scientific workflow systems and the increasing support and relevance of provenance as part of their functionality, the challenge arises to enable scientists to use provenance for the discovery of experiments, programs, and data of interest. PBase aims to satisfy this requirement while also presenting to the user a customized graphical user interface that greatly facilitates the exploration of the repository and the visualization of results.

Keywords: Provenance · Scientific workflows · Graph keyword search · Quality of service · Ranking

1 Introduction

Scientific workflow management systems (SWMSs) offer numerous advantages for computational experiments exploiting scientific data. Through a friendly user interface, the various tasks comprising an experiment can be associated with concrete computational actors (e.g. Web Services, scripts, etc.) and organized in a pipeline, which can be easily modified and shared. Furthermore, the execution environment of the SWMS often offers capabilities such as fault tolerance, distributed execution, and scalability. An additional capability of modern SWMSs is to automatically record the context and events associated with the execution of a workflow, resulting in a trace that represents the retrospective provenance of the associated data products.

Much effort has been devoted to modeling scientific workflow provenance and enabling its capture in SWMSs. In addition, graph querying techniques and their related declarative query languages have been successfully applied to provenance data, enabling its close examination for purposes such as debugging or attribution. We consider an scenario in which scientists are interested in discovering

B. Ludäscher and B. Plale (Eds.): IPAW 2014, LNCS 8628, pp. 209–214, 2015.
DOI: 10.1007/978-3-319-16462-5_17

high quality experiments, programs, and data from third parties related to their research. In this scenario, as users of a scientific workflow provenance repository, they are first likely to want to interact with the system via simple keyword searches that bring ranked results, and then possibly through a sophisticated declarative query language that yields exact results.

Therefore we introduce PBase, a scientific workflow provenance repository that enables, besides declarative queries, searching and ranking under various criteria and at different granularities. Concretely, users can search annotated workflows and traces based on criteria that apply globally to entire workflows and traces, or individually to their component actors and data products. Result items can be obtained not only if they contain the associated keywords but also if they are related to items that contain them. The criteria or facets under consideration include quality of service and authority metrics computed from the provided traces, to which additional information sources can be incorporated as well. These features are supported by a custom GUI that facilitates the visualization of workflows, their associated traces, and the search and query results.

2 Provenance-Based Searching for Scientific Workflows

PBase adopts the ProvONE[1] model which represents workflows (prospective provenance) and execution traces (retrospective provenance) in a generic manner aiming to cover the majority of SWMSs. ProvONE is serialized in an OWL 2 ontology and data instances are represented in RDF. Searching and ranking in the PBase repository is performed by keyword searches complemented by authority and quality of service criteria, which we briefly describe next.

2.1 Authority

We adopt the ObjectRank [HHP08] metric, which is applied in three variants that in turn can be combined to yield an overall ranking.

Global ObjectRank represents the overall importance of a node in a way similar to PageRank. This metric captures, for instance, that a data item that is used in important experiments may be regarded as important, whereas an ordinary experiment that uses important data may itself not be important. However, while PageRank is computed uniformly based on the links between web pages, ObjectRank is computed taking into consideration the semantics of the relations between entities. This occurs as specified by the authority transfer schema graph, which through weights established by domain experts, specifies the flow of authority across the data entities in an adjustable manner. An example authority transfer schema graph is depicted in Fig. 1 for our domain of concern.

To find entities relevant for a particular keyword query, the keyword-specific ObjectRank metric is also calculated for all keywords subject to a threshold value. In this manner nodes that do not contain the keywords but are relevant

[1] http://purl.org/provone.

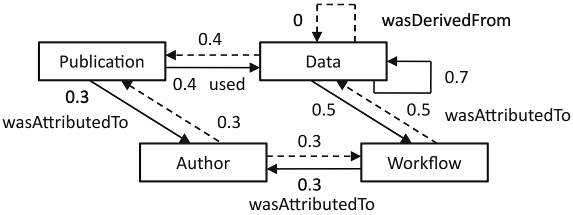


Fig. 1. Example authority transfer schema graph

for the query can be found and ranked. Finally, an inverse ObjectRank metric captures the specificity of results, placing a stronger constraint on matching the keywords of the query for cases in which the user is interested in a specific type of experiment, for example, rather than those related to a particular area. These keyword-specific metrics are computed for the most important annotated elements of both workflows and traces, i.e., actors and their executions and data items. If multiple traces are associated with a workflow, these can yield different values for a given query, due to possible missing nodes in some traces in the case of failures, for example. Collections of workflows and traces can be ranked based on the resulting values of its constituent nodes.

Furthermore, a global graph is constructed from traces to generate the ObjectRank values, whenever it is possible to identify that data generated from one workflow is used in another. This "stitching" of traces is currently limited to unique identifiers, future work involves developing alternative methods in the absence of such identifiers, by analyzing metadata, for example. If additional provenance information is available about the workflows and data, it can be incorporated into the global graph. For example, information about publications and authors as depicted in Fig. 1. Note that although the global graph is constructed from provenance information, it can be configured in various ways as required by domain experts.

2.2 Quality of Service

Numerous criteria of this type are applicable to the individual programs and data associated with workflows and their traces, the individual metrics in turn can be aggregated to assess the quality of the entire workflow, even before it is executed. In PBase we adopt the framework introduced in [CMSA04] which takes into consideration: time, cost, and reliability. The execution time is usually specified by timestamps in provenance traces, although not detailed in terms of setup and remote invocation duration, for example. We assume given measures of cost, which can be related to computational resources use or monetary cost. Reliability follows from measuring the number of times a given actor failed during its execution, which is normally inferable from traces.

The manner in which the individual metrics are aggregated for complete workflows depends on the different constructs present in the workflow as well as on the metric type. For instance, for parallel execution, the execution time of the parallel execution construct built from multiple branches corresponds to the maximum execution time of a branch. Alternatively, the reliability of a workflow built from a sequence of actors is calculated by multiplying the reliability metrics of the individual actors.

Table 1 shows the specific calculations for time, cost, and reliability (T, C, and R respectively) if we denote by c_{ij} the sequential composition of components c_i and c_j; whereas the parallel composition of a series of components c_i delimited by an *and-split* operator a and an *and-join* operator b is denoted by c_{ab}.

Table 1. Example quality of service metrics calculations

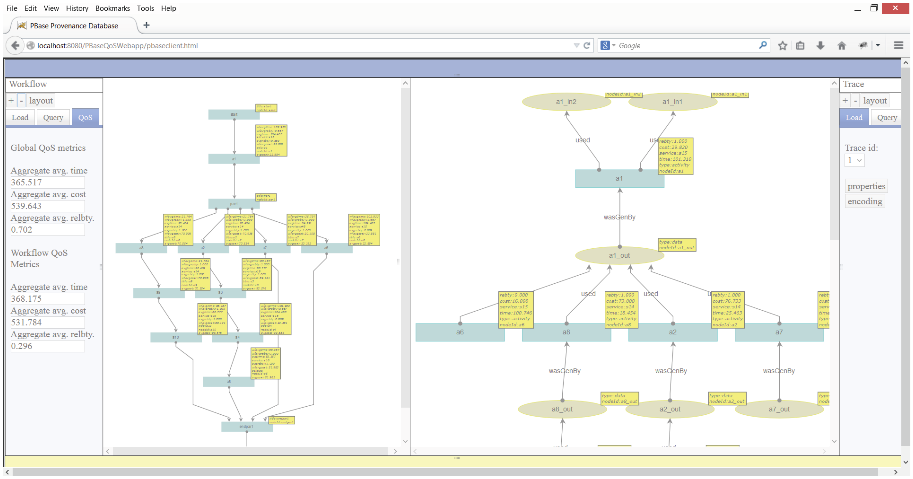
Sequential execution	Parallel execution
$T(c_{ij}) = T(c_i) + T(c_j)$	$T(c_{ab}) = max_{i \in \{1..n\}}\{T(c_i)\}$
$C(c_{ij}) = C(c_i) + C(c_j)$	$C(c_{ab}) = \sum_{i \in \{1..n\}} C(c_i)$
$R(c_{ij}) = R(c_i) * R(c_j)$	$R(c_{ab}) = \prod_{i \in \{1..n\}} R(c_i)$

3 Demonstration and Implementation

The aforementioned search criteria can be applied to the PBase repository via a GUI (see Fig. 2) that facilitates the visualization of workflows and their corresponding traces, as well as of the resulting metrics on their nodes. Keyword queries can be issued either for workflows or traces through their corresponding panels. The ranked results can be browsed over and at any time the workflow corresponding to a trace (or vice versa) can be visualized side by side. The nodes forming part of a result are also highlighted analogously and the various metrics associated with each node can be visualized by overlays next to the nodes, while global ranking lists are presented in a pop-up window.

Furthermore, it may be the case that the user is interested in a particular node, and wants to know which nodes are reachable from it (i.e. its lineage), then she can select the node and the reachable nodes are highlighted. This is done efficiently on the client side with the use of a tree cover encoding [ABJ89]. We also offer the capability to evaluate SPARQL queries on workflows and traces and visualize their results, as described in [CKL+14], which however describes an earlier version of our repository that did not include any searching and ranking functionality.

The system is implemented following a three-tier architecture, in which the user interacts with the system through a Web GUI that employs the mxGraph library for graph visualization in combination with the YUI JavaScript framework. The application logic is organized into various components that run as a Java application on the Tomcat server. Some of these components expose a series of Restful Web services that enable the interaction with the client. Communication takes place using the JSON data format. The data is stored in the TDB RDF triplestore of the Jena framework.



Global QoS metrics

Aggregate avg. time: 365.517
Aggregate avg. cost: 539.643
Aggregate avg. relbty.: 0.702

Workflow QoS Metrics

Aggregate avg. time: 368.175
Aggregate avg. cost: 531.784
Aggregate avg. relbty.: 0.296

Trace id: 1

rebty:1.000
cost:29.820
service:s15
time:101.310
type:activity
nodeId:a1

type:data
nodeId:a1_out

rebty:0.000
cost:16.008
service:s15
time:100.746
type:activity
nodeId:a6

rebty:1.000
cost:73.008
service:s14
time:18.454
type:activity
nodeId:a8

rebty:1.000
cost:76.733
service:s14
time:25.463
type:activity
nodeId:a2

type:data
nodeId:a8_out

type:data
nodeId:a2_out

Fig. 2. Graphical user interface of PBase

Currently, for testing and demonstration purposes we have created a synthetic dataset of workflows and their corresponding traces obtained via simulation, which are stored in accordance to the ProvONE model. These workflows correspond to series-parallel graphs generated randomly. Quality of service values are assigned randomly as well during the simulation process, which takes possible failures into consideration. Annotations to describe actors, data, and additional entities were obtained from myExperiment and ProgrammableWeb for example domains and are processed with the Lucene Java library.

4 Future Work

In regards to keyword relevance ranking for collections of workflows and their corresponding traces, currently we simply compute the average ranking of nodes above a certain threshold. Future work involves more sophisticated ranking techniques and exploring top-k result retrieval techniques. We also plan to incorporate quality of service metrics aggregation for various workflow models. Presently the graph algorithms run on custom Java code, future work involves exploring high performance graph libraries as well as graph distributed computing frameworks.

5 Related Work

The use of authority metrics such as ObjectRank for provenance is explored in [IHFG12], which however focuses on a generic computing framework rather than a repository. The computation of aggregate quality of service metrics has received significant attention for business processes, for example in [YDGBn+12]. Our approach aims to enable the use of techniques developed through research to enhance some of the functionality present in systems such as myExperiment [DRGS09].

Acknowledgement. This work was supported by NSF Award OCI–0830944 (DataONE).

References

[ABJ89] Agrawal, R., Borgida, A., Jagadish, H.V.: Efficient management of transitive relationships in large data and knowledge bases. In: Proceedings of the 1989 ACM SIGMOD International Conference on Management of Data, SIGMOD 1989, pp. 253–262. ACM, New York (1989)

[CKL+14] Cuevas-Vicenttín, V., Kianmajd, P., Ludäscher, B., Missier, P., Chirigati, F.S., Wei, Y., Koop, D., Dey, S.C.: The PBase scientific workflow provenance repository. Int. J. Digit. Curation **9**(2), 28–38 (2014)

[CMSA04] Cardoso, J., Miller, J., Sheth, A., Arnold, J.: Quality of service for workflows and web service processes. J. Web Seman. **1**, 281–308 (2004)

[DRGS09] De Roure, D., Goble, C., Stevens, R.: The design and realisation of the experimentmy virtual research environment for social sharing of workflows. Future Gener. Comput. Syst. **25**(5), 561–567 (2009)

[HHP08] Hristidis, V., Hwang, H., Papakonstantinou, Y.: Authority-based keyword search in databases. ACM Trans. Database Syst. **33**(1), 1:1–1:40 (2008)

[IHFG12] Ives, Z.G., Haeberlen, A., Feng, T., Gatterbauer, W.: Querying provenance for ranking and recommending. In: Proceedings of the 4th USENIX Conference on Theory and Practice of Provenance, TaPP 2012, p. 9. USENIX Association, Berkeley (2012)

[YDGBn+12] Yang, Y., Dumas, M., García-Bañuelos, L., Polyvyanyy, A., Zhang, L.: Generalized aggregate quality of service computation for composite services. J. Syst. Softw. **85**(8), 1818–1830 (2012)