Dataflow Oriented Workflow Similarity Matching Using Causal Analysis

Philip Yeo  
Faculty of Computer Science  
Dalhousie University  
6050 University Ave, Halifax, NS  
Canada B3H 4R2  
yphilip@cs.dal.ca

Syed Sibte Raza Abidi  
Faculty of Computer Science  
Dalhousie University  
6050 University Ave, Halifax, NS  
Canada B3H 4R2  
sraza@cs.dal.ca

BACKGROUND

Workflows are used to automate repetitive tasks and they resemble real-life processes. A process is made up of a collection of interrelated activities and is initiated in response to a triggering event. The goal of the process is to achieve a specific or discrete result for the requestor [2]. Workflow is an abstraction of the real process and it is represented as a model of computation (MoC) [3]. There are two types of MoC—i.e. control-driven or data-driven. In a control-driven workflow, the link is defined by a control dependency that specifies that the execution moves from activity to another while a data-driven workflow’s link is defined by the data transfer between the two activities. Most scientific workflows are dataflow oriented [1] because scientific processes are data intensive in nature and adhere to dataflow orientation as its core MoC. Workflow similarity matching is a problem that has been looked at quite extensively. However, most of these bodies of work revolve around control-flow oriented types of workflow which are associated with business workflow type. With the increase presence of scientific workflows that are mainly dataflow oriented, workflow similarity matching that targets them are equally important.

Categories and Subject Descriptors
H.3.4 [Information Storage and Retrieval]: Systems and Software – dataflow oriented workflow management system  
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – causal footprint analysis

General Terms
Algorithms, Design, Experimentation, and Theory

Keywords
Causal footprint, workflow, workflow similarity matching, workflow reuse, dataflow oriented workflow

1. OBJECTIVE

Our system, ‘BSim Match’, is used in the context of workflow modeler searching for similar workflows for the purpose of workflow reuse. It targets a set of workflow model objects given a query workflow model. This context applies in situation of preventing duplications and redundant workflow as well as to be invoked during the design stage of the workflow. The latter allows for the system to suggest similar workflow to the workflow modeler, whereby he or she can then decide to use all or part of the suggested similar workflows, thereby reducing the time taken for designing the workflow.

Currently, there is a gap in providing a similarity matching solution that allows for the workflow modeler to find similar workflow models based on a selection of matching criteria that users can specify. These criteria reflect the different features and characteristics of the workflow model that is provided by the workflow model’s specification language. The exploitation of these elements based on the purpose of the workflow model match-making is an interesting problem and one that has not been fully researched. In this work, we aim to:

- Investigate how workflow similarity matching for dataflow oriented workflows can be established using a behavioral approach and,
- Provide the ability to allow a user to select which characteristics of the workflow they intend to use for finding the similar workflows (a purpose driven approach).

Thus our objective is to introduce a behavioral analysis technique for similarity matching of dataflow oriented workflows that uses causal footprint method for discovering the workflow’s behavioral semantics.

2. APPROACH

In this paper, we demonstrate our work that applies an existing behavioral analysis technique that has been used for finding similarity in control-flow oriented workflows, against a dataflow oriented workflow scenario. We have introduced additional steps in our method that allowed us to make this possible. Furthermore, positional aspects were also used in the similarity matching process as it provided transitional information to the flow of data in its dataflow path.

Existing workflow similarity matching solutions are focused on providing the matched workflow either with a score or a Boolean answer. However, most workflow model specifications have attributes such as its workflow elements, data links, overall input and output ports that can be used leverage for finding similar workflows. This applies to dataflow oriented workflows as well. These elements are characteristics of the workflow model and can be harnessed to compare with other workflows in ascertaining its similarity. Our suggested solution does not target the run-time
phase of a workflow and is currently meant to be use in a build-time phase.

Our solution framework called ‘BSim Matcher’ for dataflow oriented workflow shown in Figure 1, comprises of five stages. Each stage contributes to the overall similarity matching process, which results in generating a score that determines the degree of similarity between two workflows provided to the system. The matching process begins with a parsing component that reads the specific workflow model specification provided and churns an internal common class object representation of the workflow model. These object classes are provided to the element term generator that then determines element terms to be detected based on the element selection provided by user. These terms are provided to the vector generator that then aligns and indexes the vectors. With the aligned and indexed vectors, calculation can be performed to determine the score between the query and the target workflow. The differentiator of our method is that it caters to dataflow-oriented workflows by using a 3-step process (T2CPF) for mapping dataflow oriented workflow to causal footprints. This method is the realization of the ‘workflow model parsing’, ‘common objects representation generator’ and ‘element term generator’ stage of the solution framework.

![BSim Matcher’s behavioral analysis approach.](image)

### 3. RESULTS

We present an experimental evaluation of each of the criteria provided by BSim Match as mentioned in Table 1 in terms of the retrieved results.

<table>
<thead>
<tr>
<th>Element Selection</th>
<th>Workflow Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor Node (PN)</td>
<td>Depending on the processor node type. For unique nodes, we use the unique location address as the label while for non-unique nodes, we use entity label</td>
</tr>
<tr>
<td>Workflow Input and Output (IO)</td>
<td>Main workflow input and output ports, and input and output parameters that are part of a data link only</td>
</tr>
<tr>
<td>Data link (DL)</td>
<td>Transitional data links (excluding data link connected to the main input and output port)</td>
</tr>
<tr>
<td>CF4DW</td>
<td>Causal footprint, ((N, DP, L_0, L_t))</td>
</tr>
</tbody>
</table>

| CF4DW-Post | Generation of dataflow paths which would aid in the creation of forward and backward links |

The experimental results have demonstrated that each of the criteria provided by our BSim Match is essential for different workflow queries modifications. Thus, a purpose-driven similarity matching system such as BSim Match provides users with enhanced workflow matching results subjective to user choice of matching criteria.

### 4. CONCLUSION

In conclusion, we have introduced a similarity matching approach for dataflow oriented workflows using a behavioral approach based on causal footprint analysis. Our final solution implementation, BSim Match, is a system that provides the capabilities to its users to select multiple criteria for matching dataflow oriented workflows. This addresses our research objectives of:

- investigating how workflow similarity matching for dataflow oriented workflows can be established using behavioural approach and
- provide users the ability to select which characteristics of the workflow they intend to use for finding the similar workflows based on.

As future line of research, we intend to experiment with and target a heterogeneous workflow environment setting. In addition, we also wish to mix and match selected criteria for the matching process. In addition to that we also plan to develop a modification strategy for the evaluation stage of the proposed project.

### 5. ACKNOWLEDGMENTS

This research is supported by a R&D grant from CANARIE, Canada through the Network Enabled Platform program.

### 6. REFERENCES


