From Scripts towards Provenance Inference

Rezwanul Huq*, Peter M.G. Apers, Andreas Wombacher
University of Twente, The Netherlands.

Yoshihide Wada, Ludovicus P. H. van Beek
Utrecht University, The Netherlands.

eScience Application

Workflow Model: Activity

- Views: data
- Interval predicates $\rightarrow$ windows
- Trigger: based on windows
- Exactly one output view

Workflow & Provenance

- **Provenance:**
derivation history of
data products starting
from its original sources.

Workflow Provenance Capture: State-of-art

Provenance-unaware Platform

Languages like Python
Tools like Excel, R

Provenance-aware Platform

Kepler, Taverna, Karma, VisTrails
STREAM, Aurora

Bridging Gap –
How?

~~Manually Building~~

Time, Training

Workflow Provenance

Problem Statement

How to capture Workflow Provenance automatically in a provenance-unaware platform?

Our Contribution: Workflow Provenance Inference

Workflow Provenance Inference: Challenge

- Capturing *data dependences* by analyzing the script.
- Translating control dependences into data dependences.

```
if DEBUG:
    value = -1
else:
    value = 1
print value
```

Which one?

Workflow Provenance Inference: Overview

¹ off-the-shelf grammar from ANTLR site

² Attributed Graph Grammar (AGG)

Provenance Graph Model

- Represented as a graph → Provenance graph

→ Windows
→ Trigger
→ Input-output ratio
→ hasOutput

Transformation Phase

- Building the initial graph
- *Preserving order between statements*
- *Maintaining versions of variables*

Transformation: An example

```
2 var1 = 100
3 var3 = var1 + 10 / 50
4 var1 = var3
```

Re-writing Phase

- A rule consists of *LHS* and *RHS*.
- A pattern matches to *LHS* will be replaced by *RHS*.
- Re-write rules for:
 - Translating *control-flow* statements
 - Maintaining persistence of views
 - Ensuring compactness of the graph

Re-writing: An example

```
11 for i in range(start, end, incr) :  
12     value = ...
```

Rule: LHS

RHS

Workflow Provenance Graph

Before Re-writing

After Re-writing

Evaluation: Use Case

- Water Scarcity Modeling.
- We focus on estimating **irrigation water demand**.
- Several files (PCRaster maps with 360*720 dimension) are used with different PCRaster operations.

Evaluation: Quantitative Analysis

Lines 120

**Initial Graph:
~ 450 nodes**

**Final Graph:
~ 139 nodes**

Workflow Provenance

**> 3000
maps**

**~ 40 GB
offline data**

**Inference
Methods**

Fine-grained Provenance Inference

Evaluation: Qualitative Analysis (I)

- Open-ended interview with two scientists
 - Debugging-friendliness
 - Extensibility
 - Customization

Evaluation: Quantitative Analysis (II)

“Sometimes, I used to spend hours finding reasons for having an unexpected value.”

Extensibility

→ Need to enter few information for the very first run.

“This is too detailed. I want to group some elements to have an overview of the processing”

Debugging-friendliness

→ Easy access to data
→ code efficiency

“I need to access library functions or functions written elsewhere.”

Customization

→ Adaptation based on user preference is possible.

Conclusion & Future Plan

- Workflow provenance capture in provenance-unaware platform
 - **Manually capturing requires both time, training**

Workflow Provenance Inference

- Future Plan
 - Address other control-flow statements
 - Build a complete framework with GUI
 - *Poster in AGU Fall Meeting 2012 at San Francisco*

Thank You ☺