NSF ACI-1443069

A Tool Supporting Collaborative Data Analytics Workflow Design and Management

PI: Jia Zhang, Carnegie Mellon University
Co-PI: Shiyong Lu, Wayne State University

Carnegie Mellon University

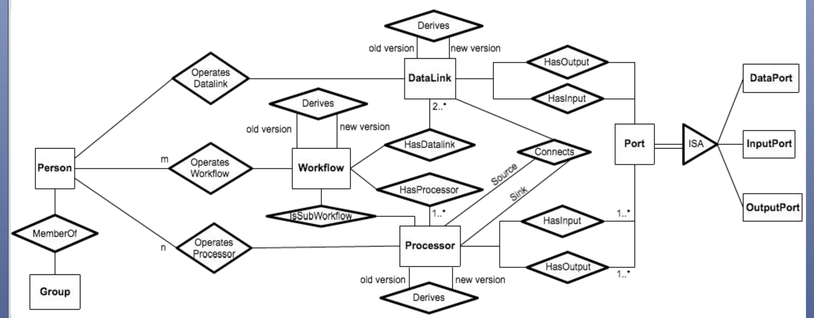
Motivation

- Big data analytics typically requires a variety of expertise that may not be realized by a single person.
- Existing scientific workflow tools are effective in supporting collaborative design.
- We aim to design and develop a technique supporting collaborative data-oriented workflow composition, as a key component toward supporting big data collaboration through the Internet.

Contributions

- Developed a collaborative provenance data model equipped with a graph-level provenance querying formalism
- Developed hypergraph theory-based algorithms for provenance management and mining
- Developed a novel software tool supporting (a)synchronous collaborative scientific workflow design, composition, reproduction, and visualization
- Extended an existing open-source workflow tool VisTrails as a proof of concept.

Collaborative Provenance Model (CPM)



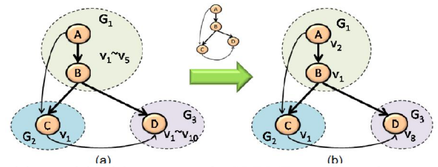
CPM CPM [@id, @name, (Add|Delete|Edit|Save)*]
Add add [@timestamp, @user, (Processor|DataPort|InputPort|OutputPort|DataLink)]
Delete delete [@timestamp, @user, (Processor|DataPort|InputPort|OutputPort|DataLink)]
Edit edit [@timestamp, @user, (Processor|DataPort|InputPort|OutputPort|DataLink), IsDerivedBy]
Save save [@timestamp, @user, Workflow, IsDerivedBy]
IsIncludedBy isIncludedBy [@timestamp, @WorkflowId|@ProcessorId|@DataLinkId]
IsDerivedBy isDerivedBy [@timestamp, @WorkflowId|@ProcessorId|@PortId|@DataLinkId]
Workflow workflow [@version, @ProcessorId|@DataLinkId, IsIncludedBy]*]
Processo processor [@id, InputPort*, OutputPort*, Function]
InputPort inputport [@id, @type, IsIncludedBy]
OutputPort outputport [@id, @type, IsIncludedBy]
DataPort dataport [@id, @type]
DataLink datalink [@id, (@InputPortId|@OutputPortId|@dataPortId, IsIncludedBy)*]

CPM-based graph patterns & graph algebra

- Extract operator (δ): extracts a set of nodes/edges from a CPM graph using a graph pattern. It takes one CPM graph (G) as input and produces a new CPM graph that matches the graph pattern as output $\delta_p(G)$.
- Union operator (∪): calculates the union of two CPM subgraphs. $G_1 \cup G_2 \rightarrow G' = (V', E')$, where:
  $V' = \{v | v \in V_1 \text{ or } v \in V_2\}$
  $E' = \{e | e \in E_1 \text{ or } e \in E_2\}$
- Intersection operator (∩): calculates the intersection of two CPM subgraphs. $G_1 \cap G_2 \rightarrow G' = (V', E')$, where :
  $V' = \{v | v \in V_1 \text{ or } v \in V_2\}$
  $E' = \{e | e \in E_1 \text{ or } e \in E_2\}$
- Difference operator (−): calculates the difference of two CPM subgraphs. $G_1 - G_2 \rightarrow G' = (V', E')$, where
  $V' = \{v | v \in V_1 \text{ and } v \notin V_2\}$
  $E' = \{e | e \in E_1 \text{ and } e \notin E_2\}$

Hypergraph Theory-based Provenance Mining

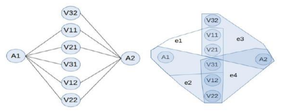
Exploratory nature of scientific workflows decides that, different versions may be equally important as they may represent different strategies and/or algorithm implementations.



Formally, we formalize CPM data as a hierarchical hypergraph, H(V, E, w, f_c), where:

- V is a set of nodes, representing various versions of an artifact and actors;
- E is a hyperedge set, $e \in E$, $e \subseteq V_1 \times V_2 \times \cdots \times V_k, k \leq |V|$. Each $e \in E$ represents an Execution Package, which contains series of different versions of artifacts that compose a workflow along with the possible actors of the workflow.
- w is a set of non-negative numbers which acts as the weight for each $e \in E$, which represents the influence of each Execution Package.
- fc: V → Bool is a consistency function that given a workflow v∈V, fc(v) will return the consistency value (true or false) of v.

- Average commute time distance similarity measure can be applied for discovery of latent associated artifacts and actors.
- Versions are modeled as hyperedges



Similarity measurement to describe relevance of nodes in hypergraph

Laplacian equation is calculated as follows:

$$L = D_{vertices} - M \times W \times D_{edge}^{-1} \times M^T$$

- D (vertices)
- D (edges)
- M: incidence matrix

$$m(v,e) = \begin{cases} 1, & v \in e \\ 0, & otherwise \end{cases}$$

- W: weight of edges

Pseudoinverse is computed as follows:

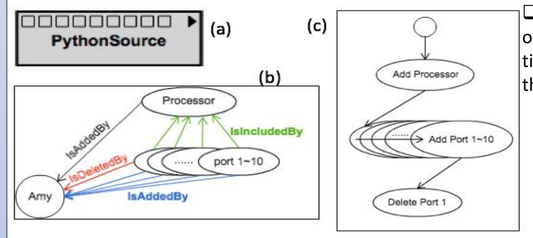
$$L^+ = (L - e \times e^T/n)^{-1} + e \times e^T/n$$

Let e_i being the *ith* column of I, the similarity is calculated:

$$S(i,j) = V\left(L^+_{ii} + L^+_{jj} - 2L^+_{ij}\right) = V(e_i - e_j)^T \times L^+ \times (e_i - e_j)$$

V = tr(D) is the volume of the hypergraph, which calculates the sum on diagonal value of D.

Prototyping System



- CPM model not only stores all operations traceable by their timestamps, but also records the structure of each entity
  - CPM can retrieve entity structure and version history easily and directly
  - Easy to find contributors to a final or immediate version of a workflow

Q1: Show all the details about how W^{v3} has been designed and evolved as it is;
match (:Person)-[r]-()-[:IsIncludedBy]->(:Entity{id:"Wv3"}) return r order by r.time*

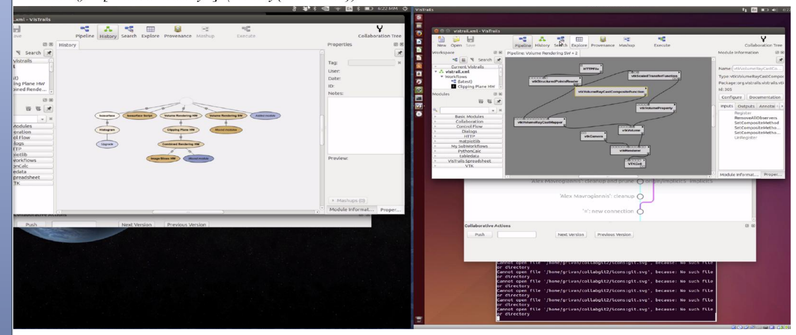
Q2: Return all the designers who contributed to the design of W^{v3};
match (:Entity{id:"Wv3"})<-[r:IsIncludedBy]-()-[:IsAddedBy|IsEditedBy|IsDeletedBy|IsMergedBy]->(p) return p*

Q3: Return the sub-workflows designed or refined by user s_1;
match ()-[r2:IsIncludedBy]-()-[r1:IsAddedBy|IsEditedBy]->(p:Person{name:"s1"}) return r1, r2

Q4: Return all user pairs who designed some workflows collaboratively;
match (p1)<-[:IsAddedBy|IsEditedBy|IsDeletedBy|IsMergedBy]-()-[:IsIncludedBy]->()<-[:IsIncludedBy*]-()-[:IsAddedBy|IsEditedBy|IsDeletedBy|IsMergedBy]->(p2) return p1,p2*

Q5: For workflow W^{v3}, which versions of comprising steps 1 and 2 are used? Who designed the two steps? How are they designed or refined? How are they merged?
match (:Person)-[r]-(e)-[:IsIncludedBy]->(:Entity{id:" Wv3"}) return e.name, e.version*

Q6: What are the previous versions of W^{v3}? Why was it refined?
match ()<-[r:IsDerivedBy]-(:Entity{id:"Wv3"}) return r*



Future Work

- Further study CPM to answer various types of queries
- Further explore hypergraph-based search algorithms
- Explore a more high-level, user-friendly language for formulating provenance queries
- Move VisTrails online

Acknowledgement

This work is sponsored by NSF ACI-1443069.

jia.zhang@sv.cmu.edu