A Data Provenance Model for Collaboration Design Process

Xuan Sun[1], Xin Gao[2*], Haiyan Kang[1], Chen Li[1]

[1]Beijing Information Science &Technology University, Beijing, China

[2] The National Computer Network Emergency Response Technical Team Coordination Center of China, Beijing, China

Keywords: data provenance; collaboration design process.

Abstract. Data provenance is the basic metadata for process analysis in process aware system and plays an important role in collaboration design process for process auditing and process improvement, but data collaboration and iterative processin theexecutionof collaboration design process impede the integrity of provenance tracking based on data provenance. This paper proposes a collaborative design provenance model based on OPM, which can support provenance tracking in the execution of multiple related collaboration design processes. In this model, it descripts the basic characteristic of the execution of collaboration design process, supports the description of iterative process by execution path expansion, and supports the descriptions of different data collaboration patternsbetween different design processes. Based on this model, we finally provide the global data provenance directed graph by the mechanism of data provenance sharing and combination that can insure the integrity of provenance tracking of collaboration design process.

1. Introduction

Collaboration design process always consist of multiple design processes, and carry out the complex design work by the multidisciplinary collaboration between different design process in the form of sharing or exchanging data. For collaboration design process, there is one key problem: how to describe the runtime state and global execution of multiple collaboration design process from the aspect of data flow. However current data provenance model focus on the process with determinate and predicable execution path of process, and they are not ready for iterative process in collaboration design process. Second, current data provenance models lack the descriptionof datacollaboration. Current data provenance models always focus on single design process and lake the description ability for the operation and data in the collaboration between multiple design processes, so that it can’t describe the evolution process of data flow among concurrent design processes. These two challenges cause that the researchers of collaboration design process can’t get the global data provenance directed graph of the execution of collaboration design process for the difficulty to insure the integrity of provenance tracking, so that they can’t use data provenance effectively in collaboration design process analysis.

To solve this problem, we try to expand current data provenance model to make it suit to describe collaboration design process. To solve this problem, we try to expand current data provenance model to make it suit to describe collaboration design process. In our preview work [1], we propose a collaboration design process model CD_net to support the description process collaboration and iterative process. Based on this work, we propose a data provenance model based on OPM [2] to describe the execution of collaboration design process based on CD_net in this paper.In this data provenance model, we provide a mechanism of execution path expansion to correspond to the description of iterative process in CD_net, which can append the data provenance informationwith finer granularity to the global data provenance directed graph. Meanwhile we provide a data provenance description mechanism of Pub/Sub data collaboration andan integration mechanism of multiple data provenance directed graphs, so that we can provide the complete data provenance information for the execution of collaboration design process.

2. Related Work

Data provenance modeling is an important part of data provenance research, and there are many researchers who have paid attention to it. In the second IPAW conference, the researchers of wright state university provide Provenir model based on W3C Web ontology language [3]. PROV-DM is a conceptual data basic model, which conforms to the standard of W3C data provenance [4]. PASOA is a data provenance model based on directed graph for process document, which can track the operational process of data and service in the SOA oriented workflow environment [5].

In the complex product design process, collaboration design is a main work pattern in this domain, and it correlated the team distributed among the world to work together for the same product design [6]. Paper [7] sets data provenance information as the basic information to check the completeness of data. Paper [8] set data provenance information as the metadata of sharing data.

3. Data provenance model for collaboration design process

The structure of CDPM is showed as Fig. 1, and it consists of six node elements and seven edge elements. The nodes of CDPM describe the correspondingentities of collaboration design process, including data, processData, collaborativeData, process, human and role.

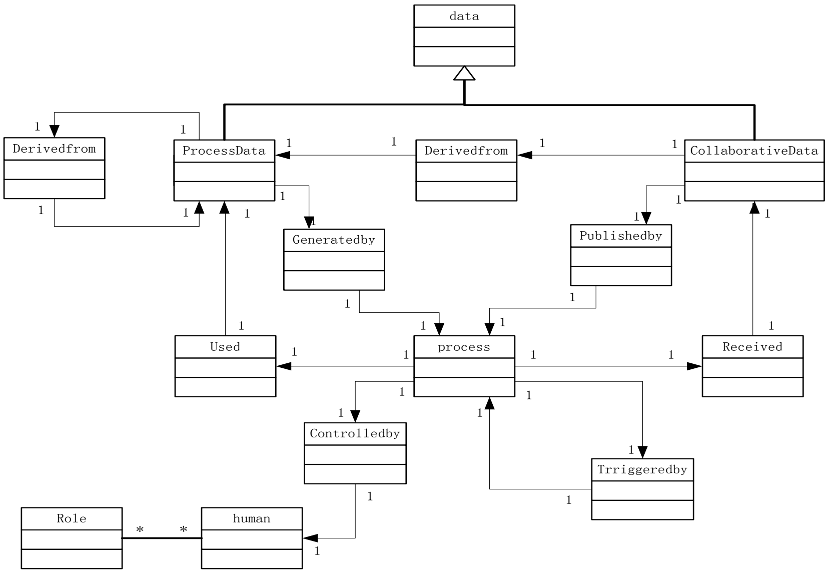

Fig.1: The structure of CDPM

- data

The data node is used to describe the intermediate data produced in the execution of collaboration design process, the reference data resource and the important design result. Because there are both the data result used for communication among the designers of the same design team and the data result shared among different design teams, the data node also has two subclasses that are “processData” node and “collaborativeData” node.

- processData

The “processData” node is mainly used to describe the data in the form of input or output for the task in the execution. It always stands for the design parameters passed from one task to the other task in the inner of design process.

- collaborativeData

The “collaborativeData” node is mainly used to describe the sharing data by the form of Publish-Subscribe among different design teams in the execution. It always stands for the design result passed from one design process to the other design process as the emergence of an instance of data collaboration among the design teams.

- process

The process node is mainly used to describe the execution instance of specific task T, and its attributes include the corresponding id, execution time, operator, operations and related constraints.

- human

The human node is mainly used to describe the operator related to specific task T, and it stands for the designer, the publisher and the subscriber among the data sharing in collaboration design process.

- role

The role node is mainly used to describe different roles for different human nodes and their corresponding function and permission. One human can control multiple process nodes, and these relations can be distinguished by the different role nodes related to the same human.

Beside the six node elements, CDPM also describes the dependency relationship between the entities in the execution instance of collaboration design process by seven edge elements which includeDerivedfrom, Generatedby, Publishedby, Used, Received, Controlledby and Triggeredby.

- Derivedfrom

"Derivedfrom" edge is the edge points from one "processData" node to another "processData" node, for example from pd2 to pd1, which means the produce of pd1 is before the production of pd2 and the state of pd1 depends on pd2. In collaboration design process, "Derivedfrom" edge always describes the transformation of specific design result, like data mapping or data filtering, and the statement of "Derivedfrom" edge betweenpd2 and pd1is as follow:

$$pd_2 \xrightarrow{\text{Derivedfrom}} pd_1 \tag{1}$$

- Generatedby

"Generatedby" edge is the edge that points from one "processData" node to one "process" node, which means the operations of "process"node need to be executed before the production of "processData" node. In collaboration design process, "Derivedfrom" edge always describes which design task produces the specific design result, and multiple "processData" nodes can point to the same "process" node by "Generatedby" edge.

- Publishedby

"Publishedby" edge is the edge that points from "collaborativeData" node to"process" node, which stands for the common data collaboration between the tasks in different design processes in the form of one-to-many data sharing. In collaboration design process, multiple "collaborativeData" nodes can point to the same "process" node by "Publishedby" edge.

- Used

"Used" edge is the edge that points from "process" node to "processData" node, which stands for the execution of "process" node needs the "processData" to be the input parameter and then "process" can carry on its operations. In collaboration design process, multiple "processData" nodes can point to the same "process" node by "Used" edge, and the execution of "process" node can start only until all the "processData" nodes are ready.

- Received

"Received" edge is the edge that points from "process" node to "collaborativeData" node, which stands for the execution of "process" node needs the "collaborativeData" to be the input parameter and then "process" can carry on its operations. In collaboration design process, multiple "collaborativeData" nodes can relate to the same "process" node by "Received" edge, and the execution of "process" node can start until all the "collaborativeData" nodes are ready.

- Controlledby

"Controlledby" edge is the edge that points from "process" node to "human" node, which stands for the execution of "process"node is controlled by the corresponding operator. Meanwhile, it always use "Role" node as the parameter of "Controlledby" edge so as to describe the duty and right of operators.

- Triggeredby

"Triggeredby" edge is the edge that points from one "process" node to another "process" node, which stands for the dependency between the instances of two tasks in the execution of collaboration design process.

4. Execution path expansion based on CDPM

We describe iterative process at runtime by provide corresponding mechanism for “Repeat” in CDPM. For these three parts of “Repeat” at runtime, we provide the corresponding description rules based on CDMP as follow:

- The input data accepted by “Repeat” is described by a “process” node called as “InputofRepeat” and a “Used” edge, where the beginning conditions of the execution of “InputofRepeat” are set as constrains of “Repeat”.
- The iterative process from T_i to T_j can be described as a set of “process” nodes. The beginning node is named as $p_{i,k}$, and the end node is named as $p_{j,k}$, where k stands for the serial number of the repeated execution of iterative process.
- The final output data of “Repeat” is described by a “process” node called as “OutputofRepeat” and a “Generatedby” edge, where the end conditions of the execution of “InputofRepeat” are set as constrains of “Repeat”.

- In the data provenance directed graph of collaboration design process, each execution of iterative process is described as a branch path from “InputofRepeat” to “OutputofRepeat”.

According these steps of path expansion, the mechanism of “Repeat” can be described as the CDPM directed graph in fig. 2.

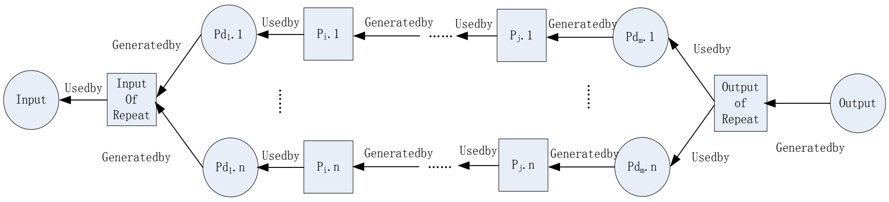

Fig.2: Path expansion of CDPM

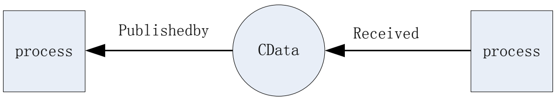

Fig.3: The Pub/Sub desription mechanisnm based on CDPM

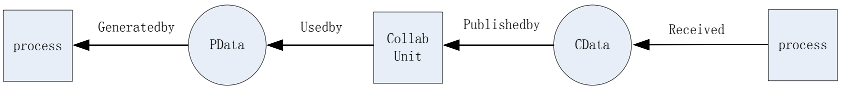

Fig.4: The refinement of Pub/Sub desription mechanisnm based on CDPM

5. Data collaboration based on CDPM

The description mechanism of Pub/Sub base on CDPM consists of “process” node, “CollaborativeData” node, “Publishedby” edge and “Received” edge, and its structure can be implementd base on CDPM as Fig. 3.

In CDPM directed graph, the node corresponding to collaboration unit in CD_net is seen as a special “process” node, which is abbreviated as “CollabUnit”. So data reference and data sharing can be described by “CollabUnit”, and the refinement of Fig. 3 can be carried out as Fig. 4.

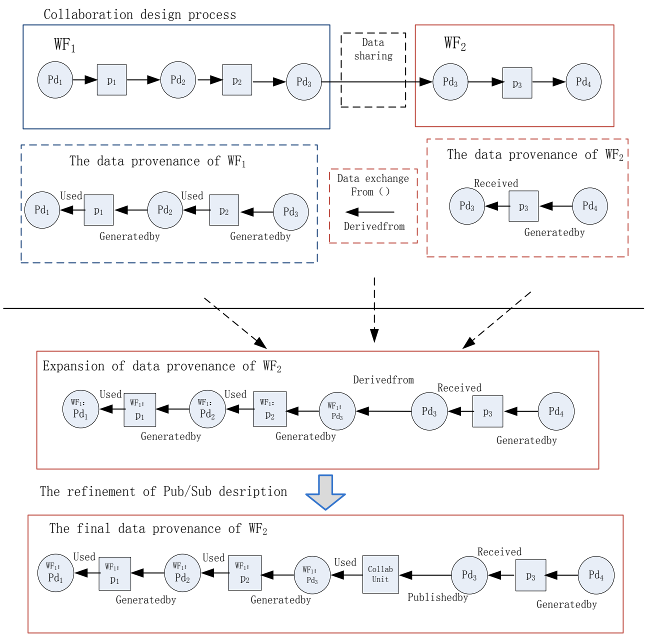

Fig.5: The combination mechinsim of CDPM deirceted graph

We need to combine the CDPM directed graphs corresponding to different design processes when they carry out data collaboration in the execution, and combination mechanism of CDPM directed graphs is shown as Fig. 5.

The steps of CDPM combination mechanism for data collaboration are as follow:

- In the stage of process design, there is data collaboration among different design processes for the design requirement of related industry. In Fig. 5, WF_1 and WF_2 are two design processes, where the data result Pd_3 is published by WF_1 and is referenced as the input of WF_2 .
- In the stage of process execution, the execution of WF_1 and WF_2 can be continued only until Pd_3 is exchanged, and the CDPM directed graphs of their execution instances are shown as the dotted boxes in Fig. 5, which is not combined. Then we use PGrap(*item*) to stand for the data provenance directed graph of specific object *item* which can be "process" node, "data" node or the execution stance of design process. Because of concurrent execution of WF_1 and WF_2 , the designer related to WF_1 only know PGrap(WF_1), but the designer related to WF_2 know PGrap(WF_2) and the origin of which is described by a "Derivedfrom" edge.
- In the stage of data Pub/Sub, the subscribers will receive and analyze the data provenanceof published data as they receive the published data, and then combine it with the data provenance of local process instance. In the process of data exchange, corresponding data provenance of the published data will be seen as a kind of metadata for published data, which is also sent to the subscribers at the same time. So the published data can be expressed by the binary group<*Dnode*, Pgrap(*Dnode*)>, where *Dnode*standsfor the data result and Pgrap(*Dnode*) stands for the corresponding data provenance directed graph. In Fig. 5, the data provenance directed graph PGrap(Pd_3) is published with the data result Pd_3 by the execution instance WF_1, and the final published data is $<Pd_3$, PGrap(Pd_3) $>$. After data exchange, the subscriber merges the received data provenance into local data provenance directed graph and then construct the complete data provenance directed graph for local proeces instance. We use UnitPGrap(WF_2) to stand for the complete data provenance directed graph of WF_2after the data provenance combination, and UnitPGrap(WF_2) can be expressed as follow:

$$\text{UnitPGrap}\ (WF_2) = \text{PGrap}(WF_2) \cup (\text{Derivedfrom:}\ (WF_2{:}Pd_3) \rightarrow (WF_1{:}Pd_3)) \cup\ \text{PGrap}(WF_1{:}Pd_3\ ) \qquad (2)$$

- After the combination of data provenance directed graphs, it still needs to be more detailed because the“Derivedfrom” edge can’t express the detail of collaboration unit in the data collaboration among different design processes. In this paper, we satisfy this requirement by the refinement of Pub/Subdescriptionmechanism based on CDPM, and the result can be expressed as follow:

$$\text{UnitPGrap }(WF_2) = \text{PGrap}(WF_2) \cup (\text{Publishedby:}(WF_2{:}Pd_3) \rightarrow \text{CollabUnit}) \cup (\text{Used：CollabUnit} \rightarrow (\ WF_1{:}Pd_3)) \cup \text{PGrap}(WF_1{:}Pd_3) \tag{3}$$

Through the combination and refinement of the data provenancedirected graphs related to both sides of data collaboration, the execution instance of single design process can be more complement, which can express the design purpose more accurate for the related designers.

6. Conclusion

In this paper, we study the data provenance model for collaboration design process which has the characteristics including iterative process and data collaboration. To get the complete data provenance directed graph for process analysis, we propose the CDPM, provide execution path expansion mechanism of CDPM to support the description of iterative process, and provide the mechanism of data provenance sharing and combination to organize the global data provenance directed graph.

Acknowledgements

Xin Gao is the corresponding author of this paper. This work was supported by a grant from the Scientific Research Funds of Beijing Information Science and Technology University(No. 1425016), National Natural Science Foundation of China under Grant No. 61272513 and National Key Technology R&D Program (No. 2012BAH45B00).

References

[1] XinGao,Wenhui Hu, Wei Ye, ZHANG Shi-kun, Xuan Sun. A data collaboration model for collaborative design based on C_net, International Conference on Software Engineering and Knowledge Engineering, pp:541-544, 2012.

[2] Moreau L，Freire J，Futrelle J，et al. The open provenance model, Southampton: School of Electronics and Computer Science，University of Southampton，2007．

[3] Sahoo S S，Barga R S，Goldstein J，et al. Provenance algebra and materialized view-based provenance management, Proc of the 2nd International Provenance and Annotation Workshop, pp: 531-540, 2008．

[4] PROV-DM: The PROV Data Model，http://www.w3.org/TR/prov-dm/.

[5] Luc Moreau, Paul Groth, et al. The Provenance of Electronic Data, Communications of the ACM, 51(4):52-58, April 2008.

[6] Hao Bo Qiu, Y. Wang, Ping Jiang, Liang Gao. Research on Workflow Modeling Methods for Collaborative Product Development, Advanced Materials Research,46:247-252, 2008.

[7] Greg Janee, James Frew and Peter Slaughter.The Data Publish Tool: A Provenance Client, http://www.alexandria.ucsb.edu/archive/2012/publish-tool.pdf.

[8] M. David Allen, Adriane Chapman, Barbara Blaustein. Provenance: Information for Shared Understanding,