ISSN:2320-0790

compusoft
An International Journal of Advanced Computer Technology

Provenance and Inputs Weigtage in a Workflow

Abhishek Narayan Shukla
School of Information and Communication Technology
Gautam Buddha University

Vimlesh Kumar
School of Information and Communication technology
Gautam Buddha University

Abstract: The provenance of data in a workflow, i.e. how the data is processed at every step and what changes made in each step. Several system are developed to capture provenance and for provenance related query which focused on result why a result is generated at a step or by what basis a particular conclusion drawn at step which data or input is ultimately responsible for error or effect on result and how much . In this paper we are focused these limitation we provide a model to provenance of data in each step of workflow and weightage of inputs on output.

Keywords: weightage; process documentation; nodes; error detection;

I. INTRODUCTION

A. *Provenance in workflow*

Now days we are focusing on error detection in workflow. We want to know how a data is processed ,in decision support and in workflow system, regulators and authority always have a concern that how a particular decision is drawn .In software application where several stages involved in a process it is highly required that each node(process) or user done their tasks accurately because the effect of each step take slight or large effect in final output of system as results in workflow are processed in several steps so single node or user do not produce or control the whole process so each process and result has to be documented . it is necessary to document each step's output data and also what change node made on previous node's output by this we can answer several provenance related query as provenance query keep concern with how instead of what so with the help of proposed model in the paper we can answer several provenance related question like-

- Which node makes the particular change C?
- Who all nodes are responsible for wrong output?
- In an output which wing or phase has max weightage?

B. *Provenance*

It is highly desirable to know Provenance of the data when data or output's authenticity has priority the provenance means to explain from where the data is derived and how [1]. Provenance increases confidence in user. It gives the surety that data or result is processed perfecty in all phased or wings of a system. Provenance also make system able to answer who is responsible for particular change or why a particular change is made

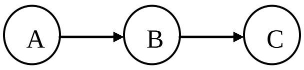


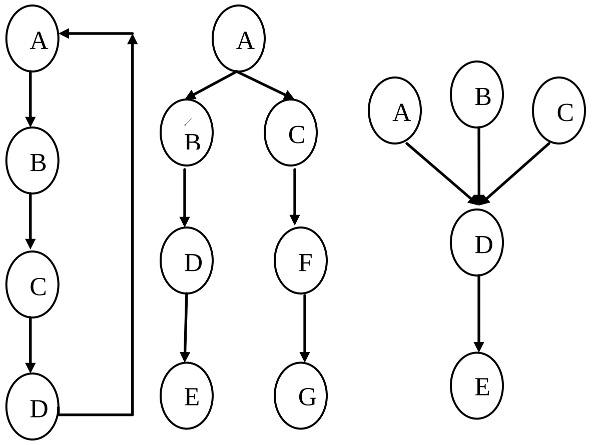
Figure 1: The Flow of Data

As an example, let us consider a workflow as shown in Figure 1 in which edges showing flow of data and vertices are process now we can say B generated an output because A gave an output and after performing some action on data received from A B generated new output which will be input for C or we can say A's output was intention which initiate the process B. Provenance also deal why a particular node or wing get interaction with other.

In provenance documentation we also have concern with why two module or wings are interacted.

C. *Workflows*

In workflow run data not choose arbitrary paths its run in same prescribed way [5]. The data in workflow may run in wings or in loop more than one wing may have same input(source generating same data for different wing) in a workflow and can generate different output. The data runs in different wings. The running data in wings merge at a point and output come after that point.



(a) Data Flow in Loop (b) Data Flow in Wings (c) Data From

Different Source

Figure 2: Different Kind of Data Flow in Workflow

Figure 1 shows linear flow of data data flow from A to B and from B to C other kind of flow is shown in Figure 2. Data flow in loop shows data flow from A to D by passing B,C,D processes and some data or all data return from D to A ,Data flow in wings same data from A flows in wing BDE and same in wing CFG ,Data from different source shown the different origins of data A,B,C which may have similar or different data sink their output to D process which give an output to E process.

II. PROCESS DOCUMENTATION

To increase authenticity of work provenance related query are good medium. To provide such provenance queries it is necessary to document the all processes. The Oxford English Dictionary defines a process as a continuous and regular action or succession of actions, taking place or carried on in a definite manner, and leading to the accomplishment of some result. So here it is necessary to document each action and we take action as node in workflow. Process Documentation model comprises of p-assertion which answer the question regarding to whole process [2]. In the example passport process several nodes interact in workflow of issuing the passport

- *Application node* is node which gets application from applicant and provides data to passport office website.
- *Website take* data from user and interface send data to data management
- *Data management node store data and provide available appointment slot also send data to verification wing and police wing.*
- *Verification wing* receive verification request from data management and verify address proof and other document.
- *Police verification wing* verifies criminal records and sends report.
- *Passport issuing authority* decides on basis of verification wing report and police wing report.

The wings also contains node but here all node are not mentioned. Between above described nodes and wings interact to each other to decide passport will be given or not. Process Documentation has p-Assertions.

A. *Interaction P -Assertion*

It document that how two node or wing get interact to each other like website may assert that they got request from user and as well data manger may assert that they get request to generate and store data from website.

B. *Node /Wing State P-Assertion*

It asserts what time request come and what time the node/wing generated output, so task completion time could be answered.

C. *Relationship P- Assertion*

The relationship p assertion tell why a node changed its state like I got this particular request so I generated that output.

III. ENGAGEMENT CHAINS

The engagement chain model presents a view of node /wing interaction as goal and intention [3]. The process documentation model has limitation that it does not documented that what was goal for which a action or process is performed. Like a person has motivation to open an account for which he applied in bank

IV. INTEGRATION OF MODELS

By integrating both model the process documentation and as well as engagement chain model for a workflow and eliminated the limitation of both models. The new integrated model for workflow contents several feature of both models like- goal interaction and intention of interaction [4]. Fr example the interaction between USER and WEBSITE as-

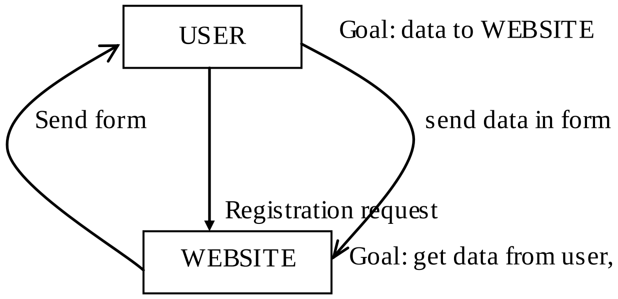


Figure 3: Integrated Model

In new model two node interacted with their own goal USER want to send data to WEBSITE where website want to receive data from user. The action performed by USER is send data in form WEBSITE action is sending form to USER.

V. LIMITATION OF INTEGRATED MODEL

Still new integrated model is not able to justify the input impact on output for example if x1,x2 are two input to a node N and output from N is O so it is un answerable in model which input have more impact on output .

VI. WEIGHETED MODEL

The new weighted model have weight when node getting input from other node the weightage can vary like single input node ‘s output fully depend on it input . if a node has two input x1,x2 and output depend on x1 more then x2 it means in big error possibility of wrong x1 is more.

In our example USER node interact with passport WEBSITE and send request to register the website send form to user and USER send data to WEBSITE. Website send this data to DATA MANGER which create 3 copy of data send one to L.I.U. one to verification officer and store one copy in database. Verification authority verify document and send report to Passport issuing authority which decide passport has to issue or not. USER Goal is registration intention is passport need action is send request to site now

WEBSITE goal is interaction between user and data manager the action is taken send form because it get request from USER send this data to DATA MANAGER because it get data from user. Now because data manager get data it send 2 copies one to VERIFICATION AUTHORITY second to L.I.U. now LIU and VERIFICATION ATHOURITY SEND REPORT TO PASSPORT ISSUING OFFICER on the basis of which he take decision . Now see the weight factor involve in output the data from website depend equally on two factor request of user and data from user both have same weightage here but in case of PIA decision is more depend on L.I.U. Report then VERIFICATION AUTHORITY. So the weightage of LIU report input is 6 and VERIFICATION ATHOURITY input is 4. If it PIA take wrong decision it means passport PIA is responsible and LIU input is wrong. VERIFICATION AUTHORITY and LIU have single input so have max weightage inputs.

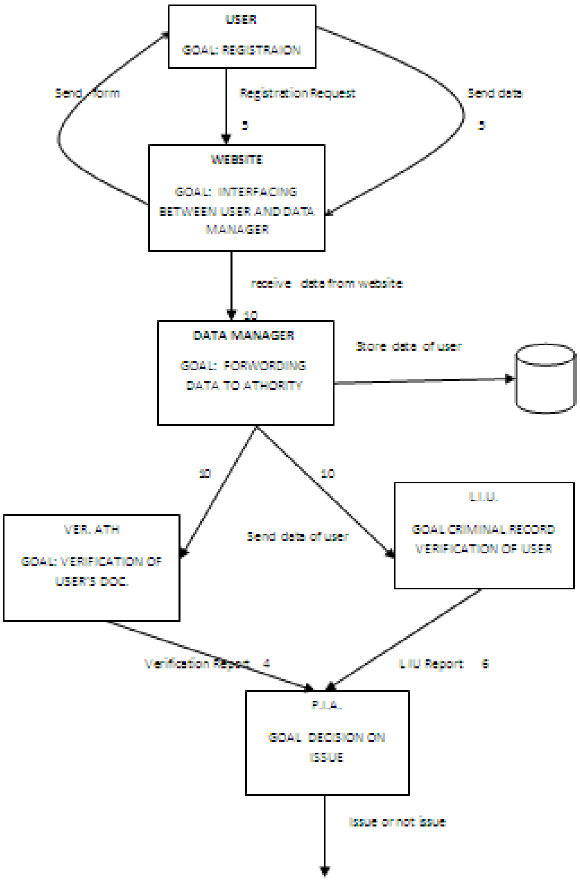


Figure 4: New weighted model

VII. ALGORITHM

In this section, we described an algorithm to find responsible node and as well as nodes input

- To find a node which is responsible for result R
- To find an interaction which is responsible for R.?

a) If R is goal of a node'N' then N is ultimately responsible for R. and if R has i1,i2,i3 input with weightage w1,w2,w3 then max weightage input will also responsilble.

b) Otherwise the node 'N' will be responsible which have an Interaction R.

VIII. EXAMPLE

We simulated this model in java application for passport process, and it gave output of responsible node. The application is developed as a project work done separately. SQL query -

1- Select * from node
Where goal = user's document verification
and document verification .

TABLE I. RESULT

N. ID	NODE	INPUT
1	LIU	Data from DATA MANAGER
2	VER ATH	Data from DATA MANAGER

IX. CONCLUSION

In this paper we presented a model which documented process as well as input with weightage. The model is able to find nodes as well as node input which are together responsible for an output while workflow are complex so its hard to understand and locate most responsible input is hard task . We hope the proposed model will be helpful in such direction it also provide transparency in decision process. An interesting direction for future work is to use the model and integrate model for expert system explanation feature as well.

ACKNOWLEDGMENT

We thank the reviewers of paper for their helpful comment and dedicated faculty members of School of ICT, Gautam Buddha University.

REFERENCES

[1] Peter Buneman,Saanjeev Khanna and Wang-Chiew Tan, "Data Provenance :Some Basic Issue",in FST TCS 2000 proceeding of the 20th Conference on foundation of software technology and theortical computer science ,2000,pp.87-93.

[2] Xian Li,Timothy Lebo,Deborah L.Mcguiness. "Provenance –Based Strategies to Develop Trust in Semantic Web Application ",in IPAW Chicaago,2006,pp 182-188.

[3] Peng Yue, Ziheng Sun, Jianya Gong, Lipping Di, "A Provenance Framework for Web Geoprocessing Workflow", in IGARSS IEEE International ,2011,pp. 3811-3814.

[4] Simon Miles, Steve Muneroe, Michael Luck,Luc Moreau, "Moelling the Provenance of Data in Autonomous Systems",in AAMAS,2007,pp.1-8.

[5] Zhuowei Bao,Susan B. Davidson,Sanjeev Khanna,Sudeepa Roy, "An Optimal Labeling Scheme for Workflow provenance Using Skelton Labels".in ACM SIGMOD,2010,pp.711-722.