On Presenting Apropos Provenance for Situation Awareness and Data Forensics

Jing Zhao, Yogesh Simmhan, and Viktor Prasanna

University of Southern California, Los Angeles, CA
{zhaoj,simmhan,prasanna}@usc.edu

Abstract. Provenance for data derived from large-scale workflows across organizations and disciplines can be complex. Users in different roles find their interpretation onerous unless it is presented in a form that is easily consumable for the given task at hand. In this position paper, we motivate the need and discuss key challenges for presenting provenance across different granularities to support data quality forensics for diverse users. We also offer potential modeling and algorithmic solutions.

1 Introduction

As data flows through and is derived from workflows executed across organizations and disciplines, provenance may be collected and reconstructed from different orchestration and execution frameworks [2], and often at different granularities depending on the execution framework in question. For example, for a workflow composed of multiple web services, the workflow management system may collect coarse-grained provenance that describes the data flow and control flow at the granularity of the web service invocations. Further, within an individual web service, detailed provenance may be collected to describe the execution logic of the service. Furthermore, more detailed provenance may be collected on system and OS calls within each execution step.

Understanding and interpreting raw provenance is challenging for users who consume it for diverse uses. The provenance collection mechanism provides a natural "grouping" structure for presenting provenance. However, it presents provenance from the perspective of the "composer" of the workflow rather than the "consumer" of the provenance. An appropriate granularity or view of provenance should be presented to users based on the current task at hand and situation of interest. For example, when using provenance for data quality debugging, fine-grained provenance needs to be provided for data objects and processes that have high impact on quality, whilst other provenance is masked. Users with different roles may also be interested in different views of provenance: business managers may be only interested in high-level business flows, while engineers are interested in detailed steps and the execution logic in the workflow.

An effective provenance presentation approach is thus required. This should determine the suitable view or granularity for provenance based on the context of usage. The presentation approach should support hybrid views that slices across vertical layers and horizontal boundaries and allow navigation across granularities. This requires support from provenance modeling, approaches to solicit

P. Groth and J. Frew (Eds.): IPAW 2012, LNCS 7525, pp. 250–253, 2012.

information on the usage context, frameworks to compose the provenance view, and presentation interfaces to display and navigate the provenance for accomplishing the task. In this paper, we outline key challenges and potential solutions for determining and presenting apropos hybrid provenance views across granularities, analyzed in the context of the Smart Power Grid domain.

2 Presenting Provenance for Data Forensics in Smart Grid

We use a use case from the Smart Power Grids domain as our motivating example. Several workflows, including the Campus Power Consumption Forecast workflow, the Forecast Model Training workflow, and the Building Sensor Integration workflow, are used to reliably forecast future power consumption of the campus, and initiate voluntary and direct-control actions to curtail energy use during peak load periods. A simplified version of the provenance collected for the ecosystem of workflows is shown in Figure 1.

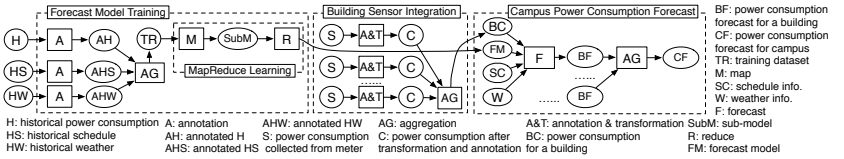
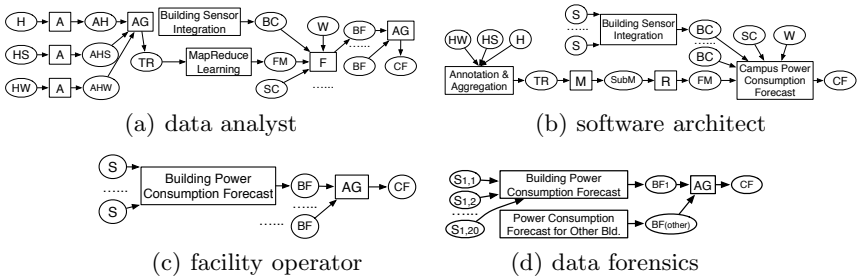


Fig. 1. Simplified provenance graph for power consumption forecast workflows



(a) data analyst (b) software architect

(c) facility operator (d) data forensics

Fig. 2. Provenance graph views for different user roles

Three types of users consume the provenance information: *the software architect*, *the data analyst*, and *the campus facility operator*. Each of them has their own particular interest and usage of provenance collected from the workflow. For example, the data analyst designs machine learning algorithms for generating the forecast model. She is interested in provenance about the execution of the forecast model training and the campus power consumption forecast workflow so that she can verify the quality of the forecast model. Figure 2 shows different provenance "views" for these user roles.

The main usage of provenance in our use case is for data quality forensics. Directly presenting a complete provenance graph with several thousand provenance nodes makes it challenging for users to perform data forensics. The *quality impact*, which indicates how the quality of a process/artifact affects the output quality, is then used to guide users on what processes and data objects they need to exercise more quality control upon. In addition to user roles, we thus also need to consider the provenance usage requirement for its presentation. Figure 2(d) illustrates a provenance view for the facility operator that reflects the granularity requirement based on quality impact. The provenance graph highlights the provenance trace for calculating the consumption forecast of Building 1 since it is the largest building and has the highest quality impact.

3 Determining Apropos Provenance Presentation View

In general, the strategies to determine the suitable provenance presentation view can be classified into a decomposition or a clustering approach. A *decomposition* approach is well suited in the presence of granularities clearly defined in the provenance model. For each individual activity in the workflow, we identify the most appropriate presentation granularity to satisfy the usage requirement and to meet the user's interest. The eventual presentation may be a combination of fine-grained and coarse-grained provenance for different sections of the graph. The approach is based on the provenance usage context information, which includes: 1) the *provenance end use* specifying the activity for which provenance is used, such as data quality forensics or software, and 2) the *user profile* describing the role of the user consuming provenance, which may include the user's affiliation, business level, associated projects, and expertise.

When existing provenance information does not have discrete granularity levels specified in the model, a *clustering* approach can be applied to infer the suitable presentation granularities. In general, this approach incrementally clusters the initial fine-grained provenance information so that groups of low-level provenance nodes are combined and replaced by new higher-level nodes. Some existing work has already discussed problems in this direction [1]. The clustering strategy needs to clearly identify what fine-grained provenance information can be combined into a composite module. A clustering strategy may also consider the semantic connection of the relevant provenance subjects. This requires mechanisms like calculation of connectivity power to be designed.

4 Conclusion

In this paper, we outlined the critical need and key challenges for determining appropriate granularities for presenting provenance. We motivated from the Smart Grid domain and illustrated alternate provenance views when presenting the same provenance to different user roles and end use needs. Our discussions centered around modeling these presentation needs and strategies to determine the appropriate view based on context information.

References

1. Biton, O., Cohen-Boulakia, S., Davidson, S., Hara, C.: Querying and managing provenance through user views in scientific workflows. In: ICDE (2008)
2. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance in e-science. SIGMOD Rec. 34 (2005)