Understanding provenance black boxes

Adriane Chapman · H.V. Jagadish

Published online: 16 January 2010
© Springer Science+Business Media, LLC 2010

Abstract Current provenance stores associated with workflow management systems (WfMSs) capture enough coarse-grained information to describe which datasets were used and which processes were run. While this information is enough to rebuild a workflow run, it is not enough to facilitate user understanding. Because the data is manipulated via a series of black boxes, it is often impossible for a human to understand what happened to the data. In this work, we highlight the missing information that can assist user understanding. Unfortunately, provenance information is already very complex and difficult for a user to comprehend, which can be exacerbated by adding the extra information needed for deeper blackbox understanding. In order to alleviate this, we develop a model of provenance answers that follow a “roll up”, “drill down” strategy. We evaluate these techniques to determine if users have better understanding of provenance information. We show how this information can be captured by workflow management systems, and that the structures and information needed for this model are a negligible addition to standard provenance stores. Finally, we implement these techniques in a real provenance system, and evaluate implementation feasibility.

Keywords Provenance · Lineage · Workflow management systems · Usability
1 Introduction

Data provenance is the history of a piece of data. This includes a description of the data’s origin, as well as the processes by which it was modified throughout its lifetime. Provenance is essential for keeping track of heavily manipulated data. Unfortunately, most provenance captured is about black-box processes, and is impossible for a user to understand.

Within a workflow management system (WfMS) that captures provenance information, the series of processes and datasets utilized is captured. Unfortunately, the information captured is often not detailed enough to answer user questions about the resulting data. In particular, we focus on extending the provenance captured about processes to facilitate greater user understanding of what happened to the data.

We provide two examples below to illustrate that the problem exists in many domains and across provenance-capturing systems. First, we look at the classic Provenance Challenge [27] workflow, and the provenance captured by the Kepler WfMS [2, 5]. We then look at a snapshot of provenance captured in MiMI [20] as our running example.

1.1 Kepler provenance example

The Provenance Challenge [27] brings several WfMSs together around a common workflow. Each system captures the provenance associated with the execution of the workflow and the results are used to make the Open Provenance Model [31] more robust. One of the key players in the challenge is the Kepler WfMS [2, 5].

Figure 1 contains a small portion of the workflow used in the Third Provenance Challenge. Figure 2 contains the Kepler provenance captured in OPM form for an small part of the workflow shown. The provenance shown in Fig. 2 contains information about the workflow nodes labeled A, B, C, D in Fig. 1 and their relationships.

1.2 MiMI provenance example

For ease of explanation, we introduce a small, but real, workflow and dataset with data items. Figure 3(a) contains the workflow used to generate the MiMI dataset [20]. Information is selected from HPRD and BIND, shown in Fig. 3(c). The resulting data items are then run through a MERGE process until no further merges take place. The workflow is shown in Fig. 3(a). We will follow three data items: LXR from HPRD and RLD and NR1H3 from BIND. The final data item is shown in Fig. 3(b). The provenance information created by executing this workflow is shown in Fig. 4.

1.3 The problem

When the workflows in Figs. 1 and 3(a) are run through a WfMS, large amounts of provenance detail is captured such as software modules used, environmental run-time

---

1We refer to the workflow management systems used in the Provenance Challenges [27], including [2, 4, 5, 9, 14, 15, 18, 26, 30, 35]. These are provenance enabled, and produce results that can be shared among systems.