Future Generation Computer Systems 36 (2014) 363–378

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Temporal representation for mining scientific data provenance

Peng Chen [a,*], Beth Plale [a], Mehmet S. Aktas [b]

[a] School of Informatics and Computing, Indiana University Bloomington, Bloomington, IN, USA
[b] Computer Engineering Department, Yildiz Technical University, Istanbul, Turkey

HIGHLIGHTS

- We propose a representation of data provenance using logical time that reduces its feature space.
- This temporal representation supports clustering, classification and association rule mining.
- Analysis of the clustering results shows that the *k*-means algorithm gives the best performance.
- We carry out an evaluation against a multi-gigabyte synthetic provenance dataset.
- We also carry out an evaluation against a real provenance dataset gathered from a satellite instrument.

ARTICLE INFO

Article history:
Received 2 February 2013
Received in revised form
2 September 2013
Accepted 5 September 2013
Available online 21 October 2013

Keywords:
Provenance
Temporal representation
Data mining

ABSTRACT

Provenance of digital scientific data is a distinct piece of metadata about a data object. It can serve as a "ground-truth" for determining the cause of execution failure for instance, or can explain a particular result to a researcher intending to reuse a data object. Provenance can quickly grow voluminous and be quite feature rich, requiring new structure and concepts that support data mining. We propose a representation of data provenance using logical time that reduces the feature space of the provenance. The temporal representation supports clustering, classification and association rule mining. This paper studies the full utility of the temporal representation through an empirical evaluation and identification of the data mining algorithms that are most effective in application to the proposed representation. The evaluation is carried out against a multi-gigabyte semi-synthetic provenance dataset built from a range of scientific workflows, and against a real one month provenance dataset gathered from a satellite instrument. Through analysis of the results via clustering metrics—purity and Normalized Mutual Information (NMI), we determine that the *k*-means algorithm gives the best clustering with the proposed temporal representation, while still yielding provenance-useful information.

1. Introduction

The provenance of a scientific data product or collection is a record of the factors contributing to the product as it exists today. That is, it identifies the what, where, when, how, and who of an object. What types of actions were applied that yielded a particular result? How and where were those actions applied? And by whom? To the extent that a data product results from raw data that itself has simple lineage, the lineage record of a data product is the latest set of activities (or "processes" in the "workflow") applied.

Provenance of digital scientific data is an important piece of the metadata of a data object. It can be used to determine attribution, to identify relationships between objects [1], to trace back differences in similar results, and in a more far reaching goal, to aid a researcher who is trying to determine whether or not an acquired data set can be reused in his or her work, by providing lineage information to support their trust in the quality of the data set. However, provenance can be highly voluminous, as capture can be carried out at a high level of granularity. This can occur for instance with a workflow system that encourages fine grained nodes (i.e., at the level of a mathematical operation) instead of coarse-grained (i.e., at the level of a large parallel computing job). The sheer volume of data has been dealt with in different ways, by developing views on the provenance [2], or by caching select content [3]. Visualization techniques are effective in making sense of large data [4]. One could throttle provenance capture to control the volume [5] of provenance generated at the source.

We take a different approach to dealing with the large volumes of provenance, and that is to assume that the volume of provenance will be large, and then selectively reduce the feature space while simultaneously preserving interesting features so that data mining on the reduced space yields provenance-useful information. The mining tasks include generating patterns that describe and

* Corresponding author. Tel.: +1 8123618295.
E-mail addresses: chenpeng@cs.indiana.edu, inspiration.chen@gmail.com (P. Chen), plale@cs.indiana.edu (B. Plale), aktas@yildiz.edu.tr (M.S. Aktas).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.09.032

distinguish the general properties of the datasets in provenance repositories (by training classifier and mining association rule set), finding variants to detect faulty provenance data (by checking cluster centroids in the case where correct and faulty provenance are naturally separated into different clusters) and discovering more descriptive knowledge of provenance clusters (by mining association rules that reflect workflow variants).

A generally accepted model for representing provenance is the Open Provenance Model (OPM) [6] which produces a directed, annotated graph of provenance entities related by causal dependencies. An OPM graph representation however is not directly useful for data mining without additional structure or structural abstraction. We propose a representation that is an abstraction of the OPM graph representation. OPM defines a historical record of dependencies between entities; hence OPM compliant graphs have implicit temporal ordering which we exploit in our proposed representation. W3C Provenance Incubator Group used OPM as the reference model in its provenance vocabulary mappings among a core set of provenance vocabularies and models [7]. One approach to extending our proposed representation to provenance in formats other than OPM is to first transform the data into OPM format using these vocabulary mappings. However, as we discuss in future work, vocabulary mappings are not the ideal solution, since they can be inaccurate and cause information loss.

We propose a representation of data provenance using logical time that reduces the feature space of the provenance. We posit that the temporal provenance representation is an efficient and useful statistical feature representation of provenance.

The goal of the study described here is to evaluate our proposed temporal provenance representation for data mining kinds of tasks. This study extends the work published in [8]. While the earlier work proposes Logical Clock-P and carries out a preliminary evaluation to show that useful data mining can be carried out against the temporal representation, the contributions of this paper are a full evaluation of the data mining potential of the Logical Clock-P representation. We carry out an empirical study to understand which clustering algorithm works the best with the proposed temporal representation. We discuss the implications of the temporal representation by illustrating the use of the representation for detecting the cause of execution failures. We evaluate the performance and the scalability of proposed method. Evaluation is carried out against a large 10 GB database of provenance traces generated from six real-life workflows [9], and a real life provenance dataset [10] captured from a ground processing pipeline of the NASA AMSR-E satellite-bound instrument.

The remainder of the paper is organized as follows: Section 2 discusses related work. Section 3 describes the datasets that we are experimenting with. Section 4 introduces the causal graph partitioning approach, while Section 5 describes the feature selection. We present our methodology of choosing mining techniques in Section 6 and show the evaluation results in Section 7. A performance evaluation of scalability of the approach appears in Section 8. Section 9 highlights important aspects of the paper and discusses future work.

2. Related work

Provenance in e-Science is first comprehensively discussed in a 2005 survey of provenance [11]. Davidson and Freire [12] provide an additional survey view of provenance. Davidson et al. [13] first introduce the problem of mining and extracting knowledge from provenance.

Margo and Smogor [14] use data mining and machine learning techniques to extract semantic information from I/O provenance gathered through the file system interface of a computer. The mining step reduces the large, singular provenance graph to a small number of per-file features. Our research is complementary in that we examine a collection of provenance graphs and treat a whole provenance graph as an entity. Like Margo's work, we also reduce the size and dimensionality of provenance, but we achieve this by partitioning the graph and applying statistical post-processing. Phala [15] uses provenance information as a new experience-based knowledge source, and utilizes the information to suggest possible completion scenarios to workflow graphs. It does not, however, provide descriptive knowledge for a large provenance dataset.

Clustering techniques have been applied to workflow graphs. A workflow script or graph is either an abstract or implementation plan of execution. A provenance graph, on the other hand, is a record of execution. A provenance record may or may not have the benefit of an accompanying workflow script, so a workflow graph is in some cases a coarse approximation of provenance graph. Santos et al. [16] apply clustering techniques to organize large collections of workflow graphs. They propose two different representations: the labeled workflow graph and the multidimensional vector. However, their representation using labeled workflow graphs becomes too large if the workflow is big, and the structural information is completely lost if using a multidimensional vector. Jung and Bae [17] propose the cluster process model represented as a weighted complete dependency graph. Similarities among graph vectors are measured based on relative frequency of each activity and transition. It has the same scalability issue as Santos et al. Our work addresses the problem of mining and discovering knowledge from provenance graphs, while overcoming the scalability issue by reducing the large provenance graph to a small temporal representation sequence, and retaining structural information together with attribute information.

There are existing works that study the workflow execution data, in particular to collect, discover and predicate workflow errors. The temporal representation that we propose is for all kinds of provenance, but we evaluate its usefulness by representing and mining workflow provenance that leads to the discovery of failed workflow executions. Thus, we share many commonalities with these works in terms of motivations and techniques. For example, Benabdelkader et al. [18] develop a software tool that collects execution information from various sources, and the information includes the error occurrence that can be used to visually explore and trace the source of errors. We use the *k-means* clustering algorithm to find centroid provenance graphs that can be further visualized to help understand the experiment and explore failures; Samak et al. [19] use clustering-based classification for early detection of failing workflows and a regression tree analysis to identify problematic resources and application job types. We also use the *k-means* clustering algorithm to separate the failed workflow executions from normal executions, but we adopt a graph matching algorithm to locate the root-causes; Silva et al. [20] present a practical method for autonomous detection and handling of operational incidents in workflow activities. They model workflow activities as Fuzzy Finite State Machines (FuSM) where degrees of membership are computed from metrics measuring long-tail effect, application efficiency, data transfer issues, and site-specific problems. While they use association rules for a predictive purpose, we use association rules for discovering variation patterns offline. However, the most important difference between our work and all others is probably that we make few assumption on the provenance dataset: we do not have the status information of workflows and their processes to tell whether they completed or failed; while clustering the provenance graphs, we do not know which workflow type the provenance graph belongs to (though we use the information of workflow type for performance evaluation). We experiment with datasets that come from the Karma [21] system that gathers structured and unstructured provenance data without the assumption of a single and coherent system.