2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks

Information Provenance for Open Distributed Collaborative System

Syed Imran Jami[1], Jemal H. Abawajy[2] and Zubair A. Shaikh[1]
[1]National University of Computer & Emerging Sciences, Karachi, Pakistan
[2]School of Engineering & IT, Deakin University, Melbourne, Australia

Abstract - **In autonomously managed distributed systems for collaboration, provenance can facilitate reuse of information that are interchanged, repetition of successful experiments, or to provide evidence for trust mechanisms that certain information existed at a certain period during collaboration. In this paper, we propose domain independent information provenance architecture for open collaborative distributed systems. The proposed system uses XML for interchanging information and RDF to track information provenance. The use of XML and RDF also ensures that information is universally acceptable even among heterogeneous nodes. Our proposed information provenance model can work on any operating systems or workflows.**

Keywords: Provenance, distributed systems, RDF, XML.

I. Introduction

Open distributed systems are frequently being used for collaborations. Examples include Grids and Cloud computing. However they mostly provide improvements in the distributed processes and services they are targeted to assist although for such collaborations they should provide mechanisms to trace all the past meaningful events and decisions taken.

Provenance is a method for tracking of the origin and history of an object that contains the log of each step in sourcing, moving, and processing the target object through which it derives the history from the source, where it originates, to its final destination [2, 3, 15, 18]. Information Provenance provides the mechanism for tracking the transformations applied on target information set by collaborators. It identifies the information about the collaborators that are involved in the information's creation and manipulation, that can be used for judgments about quality and trust [7].

Most of the current practices in provenance are limited to the tracking of data usage by focusing only on the data lineage problem [18]. These systems track data generated during simulations. The data generated is specific to the domain for which the lineage system is developed [2, 8, 19]. The current trends suggest that all computing applications that are processing data and information ranging from computational biology, high energy physics, intelligence information gathering, information and network management, healthcare; provenance has its role to play.

As the database community mainly deals with closed systems, there is a strong need of provenance research for open distributed system to provide provenance support to distributed applications. Open distributed systems such as Web Services' [10] and GRID computing [6] are common collaborative platforms. However, they lack mechanisms to trace results produced in such computing platforms. Also, many data and information management problems can be solved by tracking provenance record [2]. It helps in identifying multiple sources that produced a particular output. This makes provenance to be considered as one of the necessary requirements in Semantic Grids [6]. It can also help in tracking malicious activities performed in the Grid.

Unfortunately, most of the existing systems focus on data provenance only [1, 2, 8, 12, 14, 18]. Although for an open collaborative distributed systems, one can expect data, document or even Information from library or any online repository for sharing ideas and results. This suggests that document and information provenance system should also be developed to track the usage of information.

In this paper, we propose a domain independent provenance model, by using XML for interchanging information and RDF to track information provenance. The use of XML and RDF also ensures that information is universally acceptable even among heterogeneous nodes. Many online applications now require XML and RDF documents for providing collaboration and sharing of data since they are now becoming a standard for all the heterogeneous devices. We developed a prototype model of open collaborative information system that serves as a common platform to share the ideas and results.

978-0-7695-3908-9/09 $26.00 © 2009 IEEE
DOI 10.1109/I-SPAN.2009.112

This platform is developed as provenance aware by developing information provenance model over it. We argue that RDF can be used efficiently to record the provenance metadata as RDF Triples.

The rest of the paper is organized as follows. In Section 2, we discuss the related work in the area of provenance in general and information provenance in particular. This section also discusses the background information on provenance and open collaborative system of interest. Section 3 and 4 provides the information provenance model over collaborative system. Section 5 provides result while section 6 concludes the paper discussing the limitations of our approach and future refinement required in it.

II. Background and Related Work

The early system of provenance started to record the origin and history of a piece of data called data lineage that traced each step in sourcing, moving, and processing data in the domain of database system. Lately research in the development of provenance-aware systems has been receiving attention in the domain of simulations that are using Grids and Web Services. For example, myGrid project for Bio Informatics experiments [8] and CombeChem for chemistry experiments [19] tracks provenance for their simulation to maintain reliability of data to ensure maximum confidence. Provenance is also discussed in the context of the Service Oriented Architecture (SOA) [10, 19], which provides the underlying architectural basis for the Web Services/Grid environment.

In open distributed systems for collaboration, the domain is one of the important issues for which the system is designed for. Most of the provenance systems are designed for specialized domain and are not general in nature [18]. Domain independent systems require tracking generic datasets. Karma Framework provides a generic model for provenance by using XML in the front-end [17]. Provenance Recording for Services (PReServ) [10] proposes the model for generic data grids in heterogeneous environment. Buneman's work also provides generic model for provenance with heterogeneity that is meant for small scale distributed environment [2]. Our work differs from the existing work in that we propose a domain independent information provenance model by using XML for interchanging information and RDF to represent it. The existing systems are developed for tracking generic data only but our work is tracking the 'tagged representation of data' that ensures the tracking of any kind of digital artifact identified by URIs. The use of XML and RDF also ensures that information is universally acceptable even among heterogeneous nodes by using common ontology. This approach enables our system to record provenance metadata in open and heterogeneous system for collaboration.

Provenance recording systems are mostly application specific where APIs are embedded in workflows, operating systems or programming frameworks. [2, 3, 10, 14, 15] use these approaches in recording provenance. They however, require the change of workflow or operating system to operate. An alternative approach is employed by PASS [15] where operating system is responsible for fully automated provenance collection by refining its kernel. The Karma Framework provides a direct web service based APIs to record provenance by embedding this API with workflow engine that generates the target data [17]. Our work relieves operating systems and workflows from recording provenance information by embedding provenance recorder in mobile agents to provide automation and autonomous behavior. It also helps in running the process in heterogeneous environment since they have the embedded code with data. The HC-MAS healthcare system also integrates PReServ system with multi agents for recording provenance [13]. This model is however domain specific to health care environment.

Our model tracks Information provenance (also referred to as Knowledge Provenance in some literatures [7, 9, 11, 16]) for the open collaborative system. Knowledge Provenance Infrastructure [16] provides information provenance model for web information systems using semantic web based techniques. The work however described the model from the context of web inference engine. Our work details the working of information provenance model in collaborative environment with presenting the example of RDF logs as provenance metadata.

Figure 1 shows the architecture of our distributed collaborative system of concern to us. The figure summarizes different nodes that have JVM and Aglets deployed and are modeled in the application. Each of these heterogeneous nodes interacts with each other through Aglets that send or receive XML messages along with RDFs. JVM helps in coordinating between XML and Agents. This Web-based distributed collaborative system uses W3C based frameworks for information interchange. W3C standards are universally accepted that leads to its applicability to work in a heterogeneous environment. This enables our system in working on any environment as discussed in later subsections.

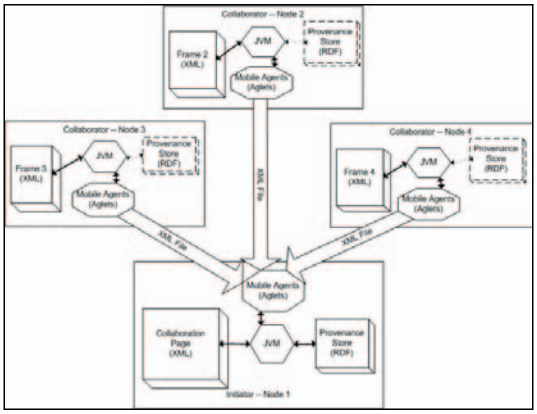

Figure 1: Distributed Collaborative Information System

The environment is implemented using Aglets for agent based communication mechanism and XML for data representation. Mobile agents using Aglets is useful in distributed applications where processing is migrated toward resources as performed in [12]. Each mobile aglet carries XML file, processing code as a class and relevant messages for processing to main user.

In the distributed collaborative system of interest, sources of information are distributed among possibly heterogeneous nodes that require open access without changing OS or programming platform. These sources may be ad hoc in nature and can lead to frequent disconnections from the system. Moreover the system must expect any of the digital artifacts that include data, document and information. To achieve these objectives, we developed open distributed system in Java to ensure platform independence while XML is used for information interchange to work in heterogeneous environment. The system in figure 1 serves as a testbed to test our provenance model. The provenance recorder is embedded in agents to provide the users with automation of recording thus relieving operating systems or workflows from doing this job. The collaborators only need to have JRE on any platform. Aglet uses JRE for execution and it is open source that can work on any platform

III. Information Provenance Metadata

In this paper an approach is proposed and developed for creating and managing information provenance metadata. We use RDF graphs to represent the provenance metadata in a form that can be understood by all the nodes in our system as shown in figure 1. In such collaborative systems information is originated by many sources, which require metadata to establish the provenance for use in many applications that are discussed previously. Such provenance metadata requires tracking the identity of the owner, the time of creation, edition or deletion and other information related to the integrity of the original document.

Information provenance metadata records all interactions between collaborators and the coordinator that initiates the collaboration. This recording helps in determining the integrity of chunks of information in an information system. The level of granularity of this system is at each frame of the XML. It can also be refined to finer level to each line depending on the requirement of application.

IV. RDF based Provenance Tracking & Representation

The information provenance metadata is represented as RDF graphs to show the relationship between information and user. Each action (insert, delete, edit, create) is represented by creating RDF sub graph which is then merged with single RDF main graph. The provenance storage boxes at each node are shown in figure 1. It shows the active RDF provenance store (solid boxes) and inactive provenance store (dashed boxes). The initiator node that is merging all RDF sub graphs from other users is storing RDF main graph in its active provenance store. The rest of the nodes' provenance store becomes active once they become initiator in collaboration.

To illustrate the working of this system we adopt this provenance aware collaborative system in which several students and their supervisors write research report collaboratively. Figure 2 shows this model in which student(s) is represented by agent A and main user – advisor is represented by agent B. This figure shows the interaction between student; which is one of the collaborators and advisor. Provenance logs as RDF triples of each interaction are recorded by agent A and merged in main RDF graph by agent B. The agents interact directly for sending XML and RDF files without requiring interference from user or application. The final RDF graph is stored along with XML document at advisor's site. The operation includes create, delete and edit each of which recorded by agent A and merged by agent B.

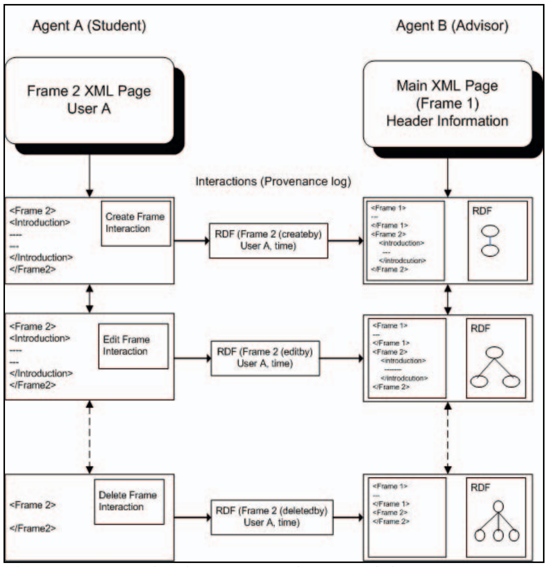

Figure 2: Interaction between Student & Advisor

The storing of these interactions as provenance information instead of standard logging systems has the advantage that complex queries can be performed over it later by other researchers. The RDF graph of final interactions allows other students and advisor to extract information related to validation of steps in a report writing process.

V. RESULTS

We now evaluate the performance of Information provenance from the aspect of recording and tracking system. Two metrics are used to evaluate provenance recording system. The major overhead in recording provenance is the size of provenance log reported which sometimes getting bigger than data itself [18]. First metric for evaluation is the size of the provenance log with respect to document size while second metric is its dependence on the frame of documents.

The evaluation is conducted on open distributed system shown in figure 1. The collaborators were running their systems under Windows XP, Linux and Mac OS X. The different nodes are connected over an Ethernet LAN connection. The systems working as nodes are Intel's Pentium 4, AMD's and Apple's Mac Machines.

The heterogeneous devices successfully reported the provenance activities in the form of RDF graphs due to incorporating mobile agents in W3C based standards. This makes our system independent of operating systems and workflows to record provenance metadata.

The results show that our provenance logs are independent of actual file size as shown in figure 3. This experiment evaluates the performance of recording provenance during 10 different collaborations. Each collaboration generates 7-10 documents over period of time. Over 3MB of textual document based data is collected that are part of our sample space. The document provenance size line in figure 3 shows the size of final RDF graph after merging that shows that the recording of provenance information is independent of actual file size thus shows the low overhead of provenance recording.

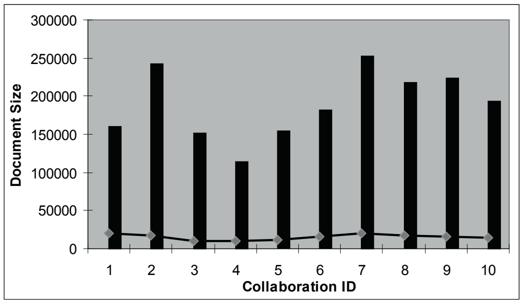

Figure 3: Provenance log dependence on file size

Our Provenance generates only one provenance log for each interaction during collaboration. This trend is shown in figure 4 that plots the bar graph of provenance log size and the number of interactions that each collaboration generates.

Each interaction is performed on different frames of the document during collaboration. The figure shows that the provenance file size after merging is highly dependent on the number of interactions. To illustrate the readings of figure 3 and 4, consider the sample of documents in collaboration ID 7 that generates highest load of document data reported in our experiment. The final document file size is close to 250KB. The final corresponding provenance log of this collaboration consisting of RDF triples is 20 KB that shows only 8% (approx.) provenance recording overhead.

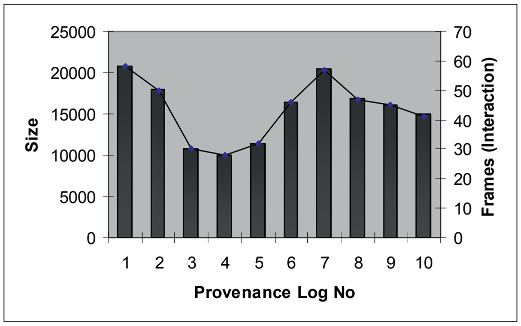

Figure 4: Relationship between provenance and frame

The total number of interactions involved in this collaboration is 57. Thus each interaction produces only one RDF graph that is finally merged with other interactions and stored at the initiator's node along with the original document.

VI. CONCLUSION

The major contribution of this paper is the proposition of information provenance architecture for an open collaborative distributed system. Information Provenance is the most neglected area in provenance. This model can be used with any Semantic Grid based system to track the usage of digital resources. The second important contribution of this paper is the recommendation and design of RDF triples to represent provenance metadata. Provenance is efficiently represented as a derivation graphs therefore RDF provides effective framework to represent the relationships. Our system is currently adopting Dublin Core Ontology for RDF provenance vocabularies which give lot of limitation in expressing RDF triples. We are currently in the process of developing RDF vocabularies for provenance ontology. The use of Aglets in our architecture achieves several benefits. It provides mechanism for automated recording of RDF graphs. However, the 'heavy' agents passing among nodes as messages result in extra communication overhead [12]. It helps in providing 'limited' autonomous nature to provenance process because this system does not rely on services from operating systems or workflows.

REFERENCES

[1]. R. S. Barga and L. A. Digiampietri, Automatic capture and efficient storage of escience experiment provenance, Concurrency and Computation: Practice and Experience, 2007. **20**(5), pp. 419-429.

[2]. P. Buneman, A. Chapman, and J. Cheney, Provenance management in curated databases, In: Proc. 2006 ACM SIGMOD International Conference on Management of Data, Chicago, USA, 2006, pp. 539-550.

[3]. A. P. Chapman, H. V. Jagadish, and P. Ramanan, Efficient Provenance Storage, In: Proc. 2008 ACM SIGMOD International Conference on Management of Data Vancouver, Canada 2008, pp. 993-1006

[6]. D. De Roure, N. R. Jennings, and N. R. Shadbolt, The semantic grid: past, present, and future, Proceedings of the IEEE, 2005. **93**(3), pp. 669-681.

[8]. C. Goble, C. Wroe, and R. Stevens, The myGrid project: services, architecture and demonstrator, In: Proc. UK e-Science All Hands Meeting, 2003.

[9]. D. P. Groth, Information provenance and the knowledge rediscovery problem, In: Proc. Eighth International Conference on Information Visualisation (ICIV 04), London, UK, 2004, pp. 345-351.

[10]. P. Groth, M. Luck, and L. Moreau, A protocol for recording provenance in service-oriented grids. Principles of Distributed Systems, Springer (2005), pp. 124-139.

[11]. J. Huang and M. S. Fox, Dynamic Knowledge Provenance. Advances in Artificial Intelligence, Springer (2004), pp. 517-523.

[12]. I. Jami and Z. Shaikh, A Multi Agent based Architecture for Data Provenance in Semantic Grid, In: Proc. International MultiConference of Engineers and Computer Scientists, Hong Kong, 2008, pp. 360-364.

[13]. T. Kifor, L. Z. Varga, J. Vázquez-Salceda, S. Álvarez, S. Willmott, S. Miles, and L. Moreau, Provenance in Agent-Mediated Healthcare Systems, IEEE Intelligent Systems, 2006. **21**(6), pp. 38-46.

[14]. G. K. Kloss and A. Schreiber, Provenance Implementation in a Scientific Simulation Environment. Provenance and Annotation of Data, Springer (2006), pp. 37-45.

[15]. K. K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. Seltzer, Provenance-aware storage systems, In: Proc. Annual Conference on USENIX '06 Annual Technical Conference Boston, MA, 2006, pp. 4-4.

[16]. P. Pinheiro da Silva, D. McGuinness, and R. McCool, Knowledge Provenance Infrastructure, IEEE Data Engineering Bulletin, 2003. **26**(4), pp. 26-32.

[17]. Y. L. Simmhan, B. Plale, and D. Gannon, Query capabilities of the Karma provenance framework, Concurrency and Computation: Practice & Experience, 2008. **20**(5), pp. 441-451.

[18]. Y. L. Simmhan, P. Beth, and G. Dennis, A survey of data provenance in e-science, ACM SIGMOD Record 2005. **34**(3), pp. 31-36.

[19]. K. R. Taylor, J. W. Essex, J. G. Frey, H. R. Mills, G. Hughes, and E. J. Zaluska, The Semantic Grid and chemistry: Experiences with CombeChem, Journal of Web Semantics, 2006. **4**(2), pp. 84-101.