A.A

Pat J

SciprovMiner: Provenance Capture using the OPM Model

Tatiane O. M. Alves, Wander Gaspar, Regina M. M. Braga, Fernanda Campos
Master's Program in Computer Science
Department of Computer Science
Federal University of Juiz de Fora, Brazil
{tatiane.ornelas, regina.braga, fernanda.campos}@ufjf.edu.br, andergaspar@gmail.com

Marco Antonio Machado, Wagner Arbex
National Center for Research in Dairy Cattle
Brazilian Enterprise for Agricultural Research - EMBRAPA
Juiz de Fora, Brazil
{arbex, machado}@cnpgl.embrapa.br

Abstract—**Provide historical scientific information to deal with loss of knowledge about scientific experiment has been the focus of several researches. However, the computational support for scientific experiment on a large scale is still incipient and is considered one of the challenges set by the Brazilian Computer Society for the period 2006 to 2016. This work aims to contribute in this area, presenting the SciProvMiner architecture, which main objective is to collect prospective and retrospective provenance of scientific experiments as well as apply data mining techniques in the collected provenance data to enable the discovery of unknown patterns in these data.**

Keywords-web services; ontology; provenance; OPM; OPMO; data mining.

I. INTRODUCTION

The large-scale computing has been widely used as a methodology for scientific research: there are several success stories in many domains, including physics, bioinformatics, engineering and geosciences [1].

Although scientific knowledge continue to be generated in a traditional way, in-vivo and in-vitro, in recent decades scientific experiments began to use computational procedures to simulate their own execution environments, giving origin to in-virtuo scientific experimentation. Moreover, even the objects and participants in an experiment can be simulated, resulting in in-silico experimentation category. These large-scale computations, which support a scientific process, are generally referred to as e-Science [1].

The work presented in this article, SciProvMiner, is part of a project developed in Federal University of Juiz de Fora (UFJF), Brazil, with emphasis on studies to building an infrastructure to support e-Science projects and has as its main goal the specification of an architecture for the collection and management of provenance data and processes in the context of scientific experiments processed through computer simulations in collaborative research environments geographically dispersed and interconnected through a computational grid. The proposed architecture must provides an interoperability layer that can interact with SWfMS (Scientific Workflows Management Systems) and aims to capture retrospective and prospective provenance information generated from scientific workflows and provides a layer that allows the use of data mining techniques to discover important patterns in provenance data.

The rest of the paper is structured as follows: In Section 2 we present the theoretical foundations that underpin this work. Section 3 describes some related work. Section 4 presents the contributions proposed in this work, the architecture of SciprovMiner is detailed, showing each of its layers and features. Finally, section 5 presents final considerations and future work.

II. CONCEPTUAL BACKGROUND

Considering scientific experimentation, data provenance can be defined as information that helps to determine the historical derivation of a data product with regards to its origins. In this context, data provenance is being considered an essential component to allow reproducibility of results, sharing and reuse of knowledge by the scientific community [2].

Considering the necessity of providing interoperability in order to exchange provenance data, there is an initiative to provide a standard to facilitate this interoperability. The Open Provenance Model (OPM) [3] defines a generic representation for provenance data. The OPM assumes that the object (digital or not) provenance can be represented by an annotated causality graph, which is a directed acyclic graph, enriched with notes captured from other information related to the execution [3].

In the OPM model, provenance graphs are composed of three types of nodes [3]:

- Artifacts: represent an immutable state data, which may have a physical existence, or have a digital representation in a computer system.
- Processes: represent actions performed or caused by artifacts, and results in new artifacts.
- Agents: represent entities acting as a catalyst of a process, enabling, facilitating, controlling, or affecting its performance.

Figure 1 adapted from [3], illustrates these three entities and their possible relationships, also called dependencies. In Figure 1, the first and second edge in the right column express that a process used an artifact and that an artifact was generated by a process. These two relationships represent data

derivation dependencies. The third edge in the right column indicates that a process was controlled by an agent. Unlike the other two mentioned edges, this is a control relationship. The first edge in the left column is used in situations where we do not know exactly which artifacts were used by a process, but we know that this process has used some artifact generated by another process. Thus, it can be said that the process is initialized by another process.

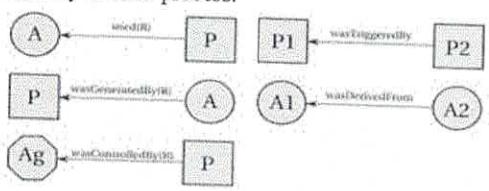


Figure 1. Edges in OPM: origins are effects and destinations are causes [adapted from [3]]

Similarly, the second edge in the left column is used in situations where we do not know what process generated a particular artifact, but we know that this artifact was derived from another artifact. These last two relationships are recursive and can be implemented considering inference rules. So, from them, it is possible to determine the sequence of execution of processes or the historical derivation that originated a data.

Some provenance models use Semantic Web technology to represent and to query provenance information. Semantic Web languages such as RDF and OWL provide a natural way to model provenance graphs and have the ability to represent complex knowledge, such as annotations and metadata[7]. One of the benefits of the semantic web approach is the ability to integrate data from any source in the knowledge base.

In [3] is defined an OWL ontology to capture concepts in OPM 1.1 version and the valid inferences in this model. Furthermore, this OWL ontology specifies an RDF serialization of the OPM abstract model. This ontology is called Open Provenance Model Ontology (OPMO).

III. Related Works

Recently, some works have been conducted to address challenges in data provenance considering scientific domain.

In [5], the author proposes an architecture for data provenance management called ProvManager. The SciProvMiner architecture as well as ProvManager is focused on capturing provenance in workflows orchestrated from heterogeneous and geographically distributed resources based on the invocation of web services. However SciProvMiner provides an infrastructure based on semantic web for provenance metadata representation and query. This feature gives to SciProvMiner a greater power of expression to infer new knowledge. ProvManager collects both prospective and retrospective provenance. SciProvMiner also collects prospective and retrospective provenance, with the differential that it uses OPM model to capture both retrospective and prospective provenance. The advantage of this approach is the interoperability of captured data, since it uses a standard model. Considering the data mining layer of SciProvMiner, ProvManager does not have.

In [4] is proposed an extension to OPM in order to model prospective provenance, in addition to retrospective provenance already supported in the native OPM model. The SciProvMiner also uses an OPM extension for supporting prospective provenance, however, the proposed architecture in [4] does not provide infrastructure based on semantic web - RDF, OWL ontologies, and inference engines, as SciProvMiner. This feature allows that the query engine of SciProvMiner can provides results that combine the power of data mining techniques together with inference engines acting over ontologies.

In [6] the authors present a mining method based on provenance data for creating and analyzing scientific workflows. SciProvMiner also uses mining methods applied to provenance data, but the focus is on unexpected knowledge discovery. Also, SciProvMiner uses ontologies together with data mining techniques.

IV. SciProvMiner- Scientific Workflow Provenance System Miner

In the context of e-Science, the emphasis on the conception and construction of SciProvMiner architecture consists of using semantic web resources to offer to researchers an environment based on interoperability for heterogeneous and distributed provenance management and query.

The SciProvMiner architecture representation considering a typical scenario is presented in Figure 2. In a collaborative environment interconnected through a computational grid, the experiments can be conducted jointly by groups of scientists located on remote research centers.

Within this scenario, researchers can model scientific workflows from different SWfMS and whose execution requires access to heterogeneous repositories and distributed in a computational grid. In this environment, the information stored lack mechanisms that contribute to a better interoperability between the data and processes generated and orchestrated within collaborative research projects. The SciProvMiner aims to provide the necessary infrastructure to make this interoperability possible.

The initial responsibility of SciProvMiner is to provide an instrumentation mechanism for the several components of the scientific workflow involved in conducting the collaborative experiment whose provenance data need to be collected for further analysis. In this context, an instrumentation mechanism is implemented using web services technology and manually configured for each component whose provenance must be collected. This mechanism aims to capture information generated during the workflow execution and send the metadata to a provenance repository managed by SciProvMiner.

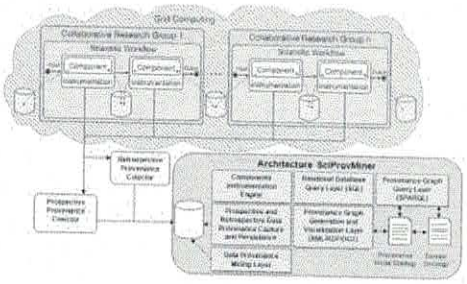


Figure 2. SciProvMiner Architecture

Considering Figure 2, "Provenance Graph Generation and Visualization Layer" has several interrelated tasks. From the relational database, the information persisted according to OPM model are processed in order to obtain an in-memory representation of the provenance graph corresponding to each execution of the scientific workflow and also of the execution models of this workflow. This layer allows the construction of a visual representation of the generated provenance graph, considering both retrospective and prospective provenance. The "Provenance Graph Query Layer" is intended to provide a mechanism and an interface for the user to formulate queries. This layer is associated with the OPM ontology and with the ontology of the studied domain (for example, dairy cattle and its parasites domain). This resource gives to SciProvMiner the possibility to process queries using inference machines that provides the capability to make deductions about SciProvMiner's knowledge bases and gain important results considering that it extracts additional information beyond those that are explicitly recorded in the generated provenance graphs.

The "Data Provenance Mining Layer" performs the search for unknown and useful patterns in provenance data using data mining techniques together with the ontologies and inference mechanisms. With this layer we aim to achieve a higher quality of provenance data, identifying possible errors and increasing the reliability of the results in the execution of an experiment.

A. *Provenance Capture*

As emphasized, the SciProvMiner presents an approach for managing provenance data independently of SWfMS. This implies that SciProvMiner is responsible for gathering of generated provenance data. An important feature of the instrumentalization model of SciProvMiner refers to the adoption of a single web service to wrap any original component of an experiment modeled from a scientific workflow. The SciProvMiner web service can be invoked one or more times for each component of the original workflow. In a typical scenario, a first instance of the web service is configured to gather the input data (artifacts, according to the OPM model) to be processed on the original component. In a second instance, it can capture information related to the original component (a process, according to the OPM model) such as start and end of the run, causal dependencies related to input and output artifacts (needed for the explicit characterization of which component used the input data and generated the output data).

It is important to observe that can be used a greater number of instrumentalization web service instances for the same original component, depending on the need to collect other entities and causal dependencies of the OPM model identified in a workflow. By hypothesis, the instrumentalization in the form adopted by this work is possible in any Scientific Workflow Management System that supports web services based on SOAP standard.

Starting from an approach of provenance collection at a process level (component), it opens the possibility of control over the data granularity to be captured. In the domain of scientific experimentation, may be important for the researcher to analyze information and prepare data lineage queries at various levels of detail. In this scenario, it becomes important that the provenance model may be able to deal with provenance information at different levels of granularity. (The OPM provides this resource in a convenient from the entity called "description" (account).

It can also be observed that different provenance descriptions relating to the same execution of a scientific workflow, represent varied lineage semantics. Therefore, by provenance data collection at different levels of granularity is possible to make analyzes and queries over the provenance data, according to the particularities of the research in progress. In this scenario, it should be noted that the strategy of provenance collection and management at various levels of abstraction can be considered as an important differential of our work allowing the parameterization of the provenance query according to the specific focus of research.

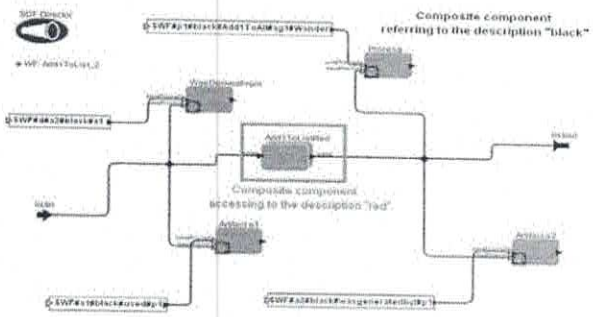


Figure 3. Composite component modeled in SWfMS Kepler

Taking an example of a workflow modeled in Kepler and the capture of retrospective provenance, shown in Figure 3, the instrumentation mechanism built manually can be used to capture the provenance data according to the OPM model in a specific level of granularity, defined according to the interests of the researcher or of the research group involved with the scientific experiment under study.

If the scientific workflow is instrumentalized in more than one level of granularity from different descriptions, the subgraph for each account (description) is displayed in a different color in the generated provenance graph.

Considering the prospective provenance context and the importance to obtain all necessary information about provenance, the SciProvMiner architecture uses an OPM

model extension as proposed in [6]. With this extension becomes possible to collect information such as the description of the workflow, the computational tasks that are part of the workflow, the subtasks of a task, the agent who performs a task and the input and output ports of a task, where the output port of a task can be connected to input port of another task, characterizing the flow of data. These informations are persisted in Relational Database Management System (RDBMS), together with information about retrospective provenance.

Due to the fact that one of the main features of SciProvMiner is the use of Semantic Web resources for representation and query of provenance data, an extension of the OPMO ontology that includes prospective provenance is currently being detailed.

V. DATA MINING

The task of data mining seeks to extract information that are hidden in large databases, which are previously unknown and potentially useful. Generally, the knowledge discovered by data mining processes is expressed in the form of rules and patterns.

According to [6], mining processes in scientific workflows are very valuable, because of scientific experiments are exploratory in nature and changes are constant.

Considering the context of SciProvMiner, data mining techniques can be applied for the following categories:

- Descriptive: whose objective is to find patterns that describe characteristics of a segment or a group of data. The main descriptive tasks are Extraction of Association Rules and Clustering, In the context of the prospective provenance data, this category is useful in order to help in the validation of workflow models and assist in the replacement of workflow models that follow similar pattern in the context of a given experiment.
- Predictive: make inferences about the existing data to build models that will be used as predictive tools. Classification and Regression may be cited as the main predictive tasks. This category is particularly interesting in the context of retrospective provenance, considering the OPMO ontology and the possibilities of connection between the data.

In this context, data mining layer of SciProvMiner acts regarding as these two categories, providing a query mechanism where data mining techniques are considered, together with OPMO and inference mechanisms. Currently, OPMO ontology is being extended to encompass specific needs of data mining techniques in order to improve data treatment.

Whereas provenance in scientific experimentation is important to help the scientist to evaluate the quality of the data, the validity of the results and allow reproducibility of the experiments, it is important to provide tools to scientist that offers more knowledge about the experiments.

VI. FUTURE WORK

In 1995, Embrapa Dairy Cattle has initiated a research project that has as main objective to obtain, by means of biological approaches, zootechnical statistics and more recently, through bioinformatics, genetic markers that, among other things, explain the existence of genetically resistant parasites and also seeking information about genetic characteristics that confer low-irritability to animals.

These studies are motivated by the importance they have and the expected production of these animals, since they are responsible for phenotypical characteristics related to animal behavior, and also the influence of this behavior on their production.

In tropical areas, the infestation of bovine animals by endo and ectoparasites causes a reduction in the productivity of animals, leading, in extreme cases, to the death of them. This issue is particularly important for Brazil, since the country has the largest commercial cattle in the world, recording in 2010 about 209 million cattles [8].

In this context, it is indispensable the collection and management of provenance data and processes of scientific experiments which will assist in the perception of historical data derivation, being necessary to the reproducibility of results obtained.

This way, the use of SciProvMiner is being inserted into this project because it is considered important to assist in the management and discovery of new patterns considering the data generated by the project.

REFERENCES

[1] S. C. Wong, S. Miles, W. Fang, P. Groth, and L. Moreau, "Provenance-based Validation of E-Science Experiments," In: 4th International Semantic Web Conference (ISWC), Galway, Ireland, 2005, pp. 801-815.

[2] J. Freire, D. Koop, E. Santos, C. T. Silva,"Provenance for Computational Tasks: A Survey", Computing in Science and Enginnering, vol. 10, n. 3, pp. 11–21, 2008.

[3] L. Moreau et al, "The Open Provenance Model core specification (version 1.1)", Future Generation Computer Systems, vol. 27 Issue 6, pp.743 –756, 2011.

[4] C. Lim, S. Lu, A. Chebotkot, F. Fotouhi,"Prospective and retrospective provenance collection in scientific workflow environments", Proceedings - 2010 IEEE 7th International Conference on Services Computing, SCC 2010, art. n. 5557202, pp. 449-456, 2010.

[5] A. Marinho et al., "Integrating Provenance Data from Distributed Workflow Systems with ProvManager", in International Provenance and Annotation Workshop - IPAW, Troy, NY, USA, pp. 0-3, 2010.

[6] R. Zeng, X. He, J. Li, Z. Liu, V.D. Aalst, "A Method to Build and Analyze Scientific Workflows from Provenance through Process Mining", 3rd USENIX Workshop on the Theory and Practice of Provenance, 2011.

[7] W. Gaspar, R. Braga, R. Campos, "SciProv: An architecture for semantic query in provenance metadata on e-Science context", 2nd International Conference on Information Technology in Bio- and Medical Informatics, ITBAM 2011, Toulouse, pp 68-81, 2011.

[8] IBGE,http://www.sidra.ibge.gov.br/bda/tabela/listabl.asp?c=73&z=t&o=24, 2011.

PROCEEDINGS

SEKE 2012

The 24th International Conference on Software Engineering & Knowledge Engineering

Sponsored by

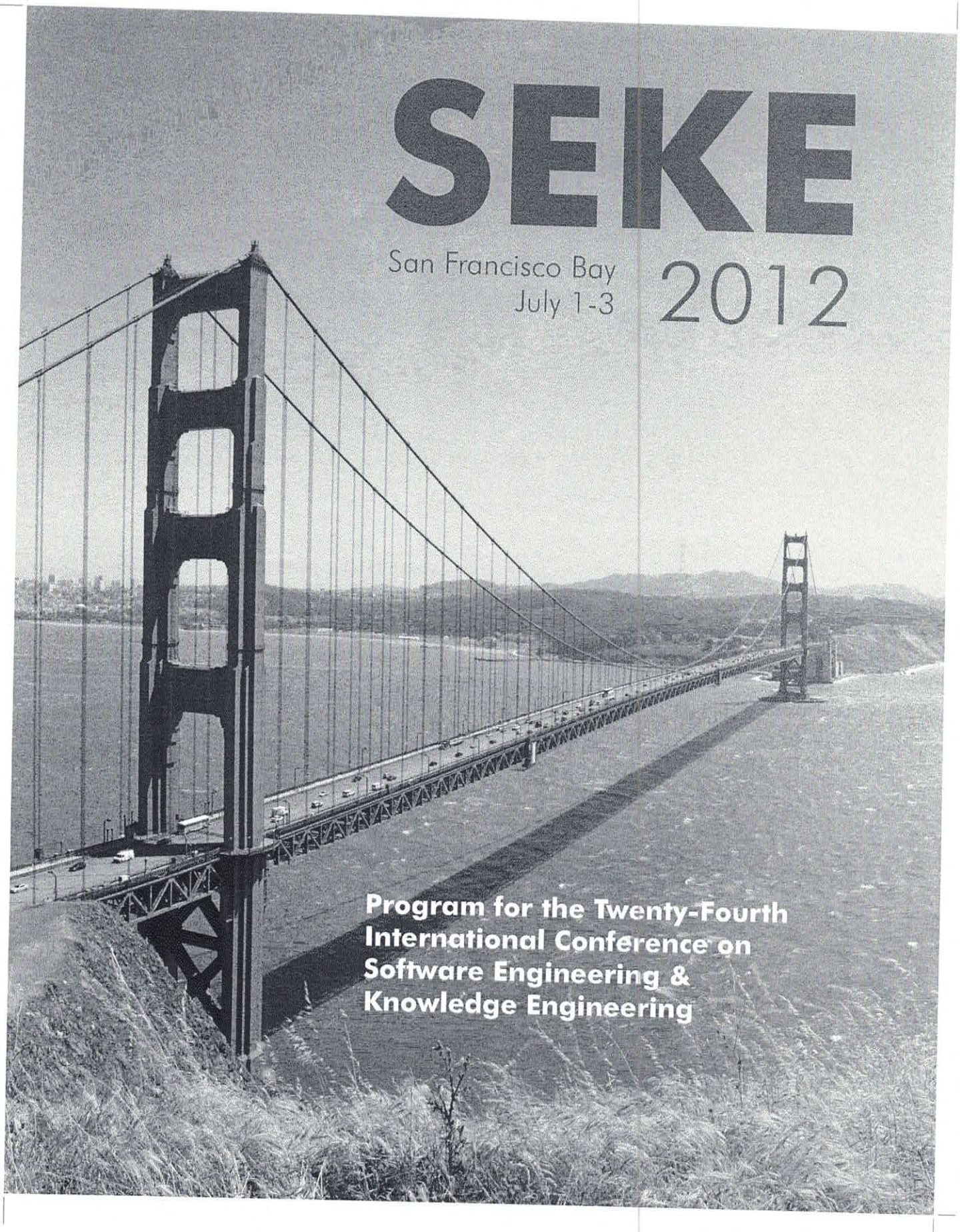
Knowledge Systems Institute Graduate School, USA

Technical Program

July 1-3, 2012
Hotel Sofitel, Redwood City, San Francisco Bay, USA

Organized by

Knowledge Systems Institute Graduate School


SEKE
San Francisco Bay
July 1-3
2012
Program for the Twenty-Fourth
International Conference on
Software Engineering &
Knowledge Engineering