香港浸會大學
HONG KONG BAPTIST UNIVERSITY

HKBU Scholars

MASTER'S THESIS

Managing service-oriented data analysis workflows using semantic web technology

Chan, Kai Kin

Date of Award:
2009

Link to publication

General rights
Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

Download date: 19 May, 2024

Managing Service-Oriented Data Analysis Workflows using Semantic Web Technology

CHAN Kai Kin

A thesis submitted in partial fulfillment of the requirements
for the degree of
Master of Philosophy

Principal Supervisor: Dr. William K. CHEUNG

Hong Kong Baptist University

September 2009

Abstract

The need to manage data analysis processes with distributed data sources and analysis services involved is getting more common nowadays for application domains like e-Science and e-Health. There exist a number of challenges still hindering from being widely adopted. In general, managing the execution of a distributed data analysis process requires a substantial overhead which includes system-level configuration, data acquisition, authentication and validation, monitoring the execution status of the current process and handling exception, comparing and reviewing different execution histories, etc. In addition, sharing the experience so that ones can leverage on each others' effort is always known to be difficult, if not impossible. Furthermore, for domains like e-Science and e-Health, the data passing along the process may contain sensitive information which implies the possibility of privacy breaching.

This thesis research aims to extend an existing workflow management system so that it can support the creation of workflow artifacts using the semantic web technology for modeling data analysis processes. First, we extend a set of ontologies for describing workflows to include concepts related to data analysis and data privacy preservation algorithms. The data analysis process artifacts created using the extended ontology do not contain any platform and dataset specific details. Thus, they can then be easily shared and reused to achieve the aforementioned knowledge sharing objective. Also, we present a rule-based policy presentation and study how enforcement of data privacy policies on data analysis workflows can readily be achieved using a rule-based inference engine. The adoption of the rule-based framework makes the system implementation to be independent to the underlying policies required, and thus is believed to be a more flexible design. Third, the

workflow artifacts are abstract in nature and cannot be executed directly unless problem specific details are added back. In addition, it needs to be recognized by the workflow execution engine. As Business Process Execution Language (BPEL) is a markup language recommended by W3C for describing web services based business processes and is supported by many existing process execution environments, we adopt it and study a particular methodology for converting a semantic workflow artifact to a BPEL process so that it can then be deployed directly it to a BPEL engine. To demonstrate the usefulness of the developed tools and methodologies, we have applied them to a hypothetic case in the context of clinical data analysis and demonstrated how one can properly manage the whole process, starting from the creation of a data analysis workflow, to policy-based workflow validation, and all the way to its final execution on a real service-oriented distributed computing environment. Some limitations and future directions of this work are also discussed in the thesis.

Keywords: Workflow, semantic web, service-oriented architecture, privacy preserving data mining, distributed data mining, clustering

Table of Contents