Chapter X
Ontology-Based Construction of Grid Data Mining Workflows

Peter Brezany
University of Vienna, Austria

Ivan Janciak
University of Vienna, Austria

A Min Tjoa
Vienna University of Technology, Austria

ABSTRACT

This chapter introduces an ontology-based framework for automated construction of complex interactive data mining workflows as a means of improving productivity of Grid-enabled data exploration systems. The authors first characterize existing manual and automated workflow composition approaches and then present their solution called GridMiner Assistant (GMA), which addresses the whole life cycle of the knowledge discovery process. GMA is specified in the OWL language and is being developed around a novel data mining ontology, which is based on concepts of industry standards like the predictive model markup language, cross industry standard process for data mining, and Java data mining API. The ontology introduces basic data mining concepts like data mining elements, tasks, services, and so forth. In addition, conceptual and implementation architectures of the framework are presented and its application to an example taken from the medical domain is illustrated. The authors hope that the further research and development of this framework can lead to productivity improvements, which can have significant impact on many real-life spheres. For example, it can be a crucial factor in achievement of scientific discoveries, optimal treatment of patients, productive decision making, cutting costs, and so forth.

INTRODUCTION

Grid computing is emerging as a key enabling infrastructure for a wide range of disciplines in science and engineering. Some of the hot topics in current Grid research include the issues associated with data mining and other analytical processes performed on large-scale data repositories integrated into the Grid. These processes are not implemented as monolithic codes. Instead, the standalone processing phases, implemented as Grid services, are combined to process data and extract knowledge patterns in various ways. They can now be viewed as complex workflows, which are highly interactive and may involve several subprocesses, such as data cleaning, data integration, data selection, modeling (applying a data mining algorithm), and postprocessing the mining results (e.g., visualization). The targeted workflows are often large, both in terms of the number of tasks in a given workflow and in terms of the total execution time. There are many possible choices concerning each process's functionality and parameters as well as the ways a process is combined into the workflow but only some combinations are valid. Moreover, users need to discover Grid resources and analytical services manually and schedule these services directly on the Grid resources essentially composing detailed workflow descriptions by hand. At present, only such a "low-productivity" working model is available to the users of the first generation data mining Grids, like **GridMiner** (Brezany et al., 2004) (a system developed by our research group), DiscoveryNet (Sairafi et al., 2003), and so forth. Productivity improvements can have significant impact on many real-life spheres, for example, it can be a crucial factor in achievement of scientific discoveries, optimal treatment of patients, productive decision making, cutting costs, and so forth. There is a stringent need for automatic or semiautomatic support for constructing valid and efficient data mining workflows on the Grid, and this (long-term) goal is associated with many research challenges.

The objective of this chapter is to present an ontology-based workflow construction framework reflecting the whole life cycle of the knowledge discovery process and explain the scientific rationale behind its design. We first introduce possible workflow composition approaches — we consider two main classes: (1) manual composition used by the current Grid data mining systems, for example, the GridMiner system, and (2) automated composition, which is addressed by our research and presented in this chapter. Then we relate these approaches to the work of others. The kernel part presents the whole framework built-up around a data mining ontology developed by us. This ontology is based on concepts reflecting the terms of several standards, namely, the predictive model markup language, cross industry process for data mining, and Java data mining API. The ontology is specified by means of OWL-S, a Web ontology language for services, and uses some concepts from Weka, a popular open source data mining toolkit. Further, conceptual and implementation architectures of the framework are discussed and illustrated by an application example taken from a medical domain. Based on the analysis of future and emerging trends and associated challenges, we discuss some future research directions followed by brief conclusions.

BACKGROUND

In the context of modern service-oriented Grid architectures, the data mining workflow can be seen as a collection of Grid services that are processed on distributed resources in a well-defined order to accomplish a larger and sophisticated data exploration goal. At the highest level, functions of Grid workflow management systems could be characterized into build-time functions and run-time functions. The build-time functions are