Geophysical Research Abstracts
Vol. 12, EGU2010-13967, 2010
EGU General Assembly 2010

ADMIRE framework for data mining and integration

Ladislav Hluchy, Viet Tran, and Ondrej Habala

Institute of Informatics, SAS, Department of Parallel and Distributed Computing, Bratislava, Slovakia (viet.ui@savba.sk)

In this paper we presents the data mining and integration of environmental applications in EU IST project ADMIRE. It briefly presents the project ADMIRE and data mining of spatio-temporal data in general. The application, originally targeting flood simulation and prediction is now being extended into the broader context of environmental studies. We describe several interesting scenarios, in which data mining and integration of distributed environmental data can improve our knowledge of the relations between various hydro-meteorological variables.

The project ADMIRE aims to deliver a consistent and easy-to-use technology for extracting information and knowledge. Its main target is to provide advanced data mining and integration techniques for distributed environment. In this paper, we will focus on one of its pilot applications with target domain is environmental risk management. Several scenarios have been proposed including short-term weather forecasting using radar images, complex hydrological scenarios with waterworks, measured data from water stations and meteorological data from models. Historical data for mining are supplied mainly by Slovak Hydrometeorological Institute and Slovak Water Enterprise. The main characteristics of data sets describing phenomena from environment applications are spatial and temporal dimensions. Integration of spatio-temporal data from different sources is a challenging task due to those dimensions. Different spatio-temporal data sets contain data at different resolutions and frequencies. This heterogeneity is the principal challenge of geo-spatial and temporal data sets integration – the integrated data set should hold homogeneous data of the same resolution and frequency.

In alignment with ADMIRE project vision, the processing elements (a data integration workflow that can be executed at a single resource.) are specified in Data Mining and Integration Language DISPEL that is being developed within the project. The goal of DISPEL is to be a canonical representation of data integration process, described in an implementation independent manner. The instance of processing elements is specified by the DISPEL description of the process that should be executed, the specification of the data/processing resource it should be executed at and instance identifier that is unique within the workflow specification.

The prototype of proposed data integration engine for environmental data is implemented in JAVA programming language. It uses OGSA-DAI framework as the platform for exposing data resources in the distributed testbed and for executing the partial workflows of processing elements; it also provide us with the data transfer capabilities and streaming of the list of tuples between remote nodes. The data integration engine takes as inputs the integration task parameters and workflow specification in DISPEL. From the workflow specification, the engine constructs an oriented graph of processing elements (defined by the mapping between inputs and outputs of processing elements), compiles the workflow to OGSA-DAI workflow objects, submits the workflow to OGSA-DAI service for execution, monitor the execution status and retrieve the result of data integration.

The graphical editor for workflows is in progress and should allow experts in environmental applications to make data integration easily. This work is partially supported also by VEGA project 2009-2011 and APVV project DMM. URL : http://www.admire-project.eu/