Future Generation Computer Systems 26 (2010) 87–96

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

A grid portal for solving geoscience problems using distributed knowledge discovery services

Gianluigi Folino, Agostino Forestiero, Giuseppe Papuzzo, Giandomenico Spezzano *

Institute of High-Performance Computing and Networking (ICAR), National Research Council (CNR), Via Pietro Bucci 41C, I-87036 Rende (CS), Italy

ARTICLE INFO

Article history:
Received 23 December 2007
Received in revised form
29 July 2009
Accepted 3 August 2009
Available online 6 August 2009

Keywords:
Grid computing
Geoscience applications
Grid services
Knowledge discovery services
Spatial data mining
Swarm intelligence

ABSTRACT

This paper describes our research effort to employ Grid technologies to enable the development of geoscience applications by integrating workflow technologies with data mining resources and a portal framework in unique work environment called MOSÈ. Using MOSÈ, a user can easily compose and execute geo-workflows for analyzing and managing natural disasters such as landslides, earthquakes, floods, wildfires, etc. MOSÈ is designed to be applicable both for the implementation of response strategies when emergencies occur and for disaster prevention. It takes advantage of the standardized resource access and workflow support for loosely coupled software components provided by web/grid services technologies. The integration of workflows with data mining services significantly improves data analysis. Geospatial data management and mining are critical areas of modern-day geosciences research. An important challenge for geospatial information mining is the distributed nature of the data. MOSÈ provides knowledge discovery services based on the WEKA data mining library and novel distributed data mining algorithms for spatial data analysis. A P2P bio-inspired algorithm for distributed spatial clustering as an example of distributed knowledge discovery service for intensive data analysis is presented. A real case application for the analysis of landslide hazard areas in the Campania Region near the Sarno area shows the advantages of using the portal.

1. Introduction

Natural disasters, such as volcanic eruptions, earthquakes, landslides, hurricanes, floods, wildfires, droughts, and tornadoes are complex physical phenomena that may cause extensive damage to property and the economy and pose a significant risk of loss of human life. Grid technologies provide a unifying infrastructure to share capabilities, to integrate services and resources, and to develop active collaborations across distributed, multi-organizational environments. Grids can be considered to be a crucial enabling technology for disaster prevention and the implementation of response strategies when emergencies do occur. Grids can help the decision making process for natural disasters such as floods, landslides and earthquakes, predicting their spread and progression as accurately as possible. The ability of Grids to address effectively emergency operations is that to combine into a single cooperating system wireless sensor networks, geospatial data, modeling and simulation capabilities, parameter estimation methods, spatial data mining tools and visualization applications. For example, Grids can coordinate computations composed of a tridimensional elevation model combined with simulation of earthquakes, to assist with disaster prevention and mitigation by creating hazard map for evacuation planning. A similar thing can be done for landslides, combining high-resolution elevation data with large-scale computer simulation to create an early warning system for evacuation.

Geospatial information is essential for quick and effective response during disaster and emergency situations. Typically, geospatial data contain a huge amount of geometric or topological information, maps and repositories of remote-sensing images. Data is often inherently distributed into several datasets, making a centralized processing of this data very inefficient. Data mining techniques can be applied to geospatial data to extract knowledge for geoscience applications. For example, the spatial prediction of landslide hazards is one important field of geoscientific research in which classification rules have been successfully applied [1]. The aim of these methods is to identify areas that are susceptible to future landsliding, based on the knowledge of past landslide events and terrain parameters, geological attributes, etc. and possibly considering anthropogenic environmental conditions associated with the presence or absence of such phenomena. As geospatial datasets are large and the data mining tasks to perform are quite complex, it is important to define novel efficient data mining algorithms that are suitable to discover meaningful patterns in order to answer scientific questions and facilitate the understanding of phenomenon.

* Corresponding author.
E-mail addresses: folino@icar.cnr.it (G. Folino), forestiero@icar.cnr.it (A. Forestiero), papuzzo@icar.cnr.it (G. Papuzzo), spezzano@icar.cnr.it (G. Spezzano).

0167-739X/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2009.08.002

Speeding up the execution of data mining tasks, and scaling the algorithms to run for large datasets is one of the most important problems in data mining research.

This paper presents the MOSÈ[1] (*Spatio-Temporal **MO**delling of Environmental Evolutionary Processes by means of Geo**SE**rvices*) system, a Grid based problem solving environment (PSE) for the developing of geoscience applications. MOSÈ is a PSE able to support the activities that concern the modeling and simulation and mining of spatio-temporal phenomena for analyzing and managing the identification and the mitigation of natural disasters like floods, wildfires, landslides etc. The activities managed by MOSÈ are characterized by the need to handle large amounts of spatio-temporal data and to support the interoperability among simulation models, distributed GIS, visualization systems, parameter estimation services, discovery of spatio-temporal patterns in pre-existing data, etc. In this domain, the data conversion and the access, search, discovery and organization processes are complex problems because data are geo-referenced, stored in distributed GIS and can be used along three dimensions: temporal, spatial and referred to the physical properties.

MOSÈ uses a Grid service based computing portal architecture to coordinate the access to the resources. Workflow technology is used to compose the services. The main components of MOSÈ are simulation services, geographic information (GI) services, knowledge discovery services (KDS), visualization services, geographic data and repositories. MOSÈ enables the creation, execution and monitoring of geo-workflows in grid environments through high level, graphical Web interfaces. Components of the workflows can be sequential, parallel and P2P applications. Each component is wrapped as Web/Grid Service for exploiting the potentialities of this architecture. Each Web service is semantically annotated and, consequently, domain specific ontologies support the user in building complex workflows, even without a deep knowledge of the domain itself.

MOSÈ provides web based access to the spatial data by a browser and allows data to be observed and manipulated in a 2D/3D space by selecting regions in thematic maps. Natural phenomena can be modeled by cellular automata (CA) models and simulated by a parallel Grid service based on the CAMELotGrid environment [2]. MOSÈ provides KDS based on the WEKA (Waikato Environment for Knowledge Analysis) data mining library and novel distributed data mining algorithms for spatial data analysis. Distributed data intensive mining algorithms are necessary to discover spatial patterns from large geospatial datasets. Novel algorithms must be developed to accomplish this task efficiently. We present an example of innovative KDS based on a bio-inspired P2P agent based algorithm for clustering distributed intensive geospatial data. The algorithm was implemented using the JXTA platform and then wrapped as a Web Service and integrated in the MOSÈ environment.

A first prototype of MOSÈ, available at the URL http://www.icar.cnr.it/mose, was successfully applied for the analysis of landslide hazard areas in the Campania Region near the Sarno area [3]. In this scenario, the main actor is a disaster manager who wants to get an overview of the Sarno area with the indication of the regions which are currently slid down and those which are susceptible to sliding down (landslide hazard areas) within a fixed time. For each scenario, the disaster manager generates a geo-workflow that orchestrates the web services necessary to obtain the outcome, and submitted it to the MOSÈ workflow enactment engine, which takes care of its execution.

Note that some of the components that constitute the MOSÈ system use results of previous research developed in the past years and guarantee high performance and accuracy of the results [4].

The paper is organized as follows: Section 2 presents the MOSÈ system and its architecture; Section 3 shows an example of the building of a workflow; Section 4 gives an overview of the distributed knowledge service approach; Section 5 illustrates a distributed knowledge service that enables the mining of geographically dispersed sites by means of a distributed multi-agent spatial clustering algorithm; Section 6 shows a real application scenario in which the distributed knowledge service and other grid services are combined to carry out an application for the analysis of landslide hazard areas in the Campania Region. Section 7 reviews some related works; finally, conclusions are drawn in Section 8.

2. MOSÈ: A Grid Portal for geoscience applications

MOSÈ is a framework supporting the development, execution and management of complex geo-models. It provides a friendly environment that symbiotically combine computations, experiments, observations and geospatial data and provide important insights into complex phenomena. MOSÈ can deliver complex grid based solutions of geoscience problems to users wherever they have access to a web browser running on the Internet without the need to download or install specialized software or worry about setting up networks, firewall, and port policies.

The framework provides a very flexible service-oriented programming environment to design and construct large scale and computationally intensive geoscience applications based on existing web/grid services technologies and standards. Remote services are visible to the users through a Web browser portal that is used to configure, launch and monitoring complex geo-workflows that combine services that communicate one another via well-defined interface and protocols. In MOSÈ the geo-workflows are tools for designing and conducting computational experiments. Scientists need to be able to run analysis processes on collected data. Often these analysis processes are single computations and often they are complex composed scenarios of preprocessing, analysis, post-processing and visualization. Furthermore, these experimental geo-workflows are often repeated hundreds of times with slightly different parameter settings or input data.

A critical feature of any e-Science portal is the capability to compose workflows, to add new computational analysis programs to a catalog of workflow components and a simple way to run the workflows with the results automatically stored in the user's private data space. MOSÈ provides a Web GUI to support easy problem set-up and data input using service classification by domain-dependent ontological annotation.

The MOSÈ Grid portal employs a service-oriented architecture and is built on the layer of OGSA (*Open Grid Services Architecture*) middleware.

The architecture, shown in Fig. 1, includes some components exported as web/grid services, each with an associated repository preserving historical (or previously inserted) information, a workflow executor and Web based access to a Geographical Information System (GIS). These services are accessed and composed by the users though the Grid Portal that contains an interface to configure, launch and monitor complex service-oriented applications such as geo-workflows for conducting experiments or other scientific investigations.

Geo-workflows greatly simplify the process of conducting geophysical analyses and forecasts. In MOSÈ, the experiments are organized as a process that represents the automation of a sequence of interactions with a set of Web Services. New experiments can be based on existing geo-workflows or created from scratch. While the workflow runs, it notifies the user of the status.

[1] http://www.icar.cnr.it/mose.