Neurocomputing 122 (2013) 88–99

Contents lists available at ScienceDirect

Neurocomputing

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

Intelligent grid enabled services for neuroimaging analysis

Richard McClatchey [a], Irfan Habib [a], Ashiq Anjum [b,*], Kamran Munir [a], Andrew Branson [a], Peter Bloodsworth [c], Saad Liaquat Kiani [a], and the neuGRID Consortium

[a] Centre for Complex Cooperative Systems, University of the West of England, UWE, Bristol B1QY, UK
[b] School of Computing and Mathematics, University of Derby, Derby DE221GB, UK
[c] School of Electrical Engineering & Computer Science, National University of Science and Technology (NUST), Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 3 October 2012
Received in revised form
3 January 2013
Accepted 5 January 2013
Available online 4 July 2013

Keywords:
Intelligent services
Machine learning and genetic algorithms
Grid enabled planning and execution
Service oriented architecture
Neuroimaging analysis

ABSTRACT

This paper reports our work in the context of the neuGRID project in the development of intelligent services for a robust and efficient Neuroimaging analysis environment. neuGRID is an EC-funded project driven by the needs of the Alzheimer's disease research community that aims to facilitate the collection and archiving of large amounts of imaging data coupled with a set of services and algorithms. By taking Alzheimer's disease as an exemplar, the neuGRID project has developed a set of intelligent services and a Grid infrastructure to enable the European neuroscience community to carry out research required for the study of degenerative brain diseases. We have investigated the use of machine learning approaches, especially evolutionary multi-objective meta-heuristics for optimising scientific analysis on distributed infrastructures. The salient features of the services and the functionality of a planning and execution architecture based on an evolutionary multi-objective meta-heuristics to achieve analysis efficiency are presented. We also describe implementation details of the services that will form an intelligent analysis environment and present results on the optimisation that has been achieved as a result of this investigation.

1. Introduction

Alzheimer's disease is a progressive, degenerative and irreversible brain disorder that causes intellectual impairment, disorientation and eventual death. It is the most common cause of dementia, accounting for around two thirds of cases in the elderly. It is estimated that 2%–5% of people over 65 years of age and up to 20% of those over 85 years of age suffer from the disease. The study of Alzheimer's disease (AD), its causes, its symptoms and especially its early diagnosis is now a major driver in the provision of healthcare for the elderly. Early diagnosis is beneficial for several reasons. Having an early diagnosis and starting treatment in the early stages of the disease can help preserve function for months to years and can aid caring strategies and support networks.

Distributed computing infrastructure based workflows are being utilised in a wide range of scientific research domains [1,2]. Alzheimer's clinical researchers are currently seeking the assistance of large-scale information technology resources to enable them to study masses of neuroimaging data being accumulated across the older patient community so that early onset indicators such as cortical thinning can be studied [3,4]. Rapid advances in neuroimaging technologies such as PET, SPECT, MR spectroscopy, DTI and fMRI have offered a new vision into the pathophysiology of AD [5] and, consequently, new increasingly powerful data analysis methods have been developed [6]. Since the beginning of the new century the development of innovative techniques for ROI-based volumetry, automated voxel based morphometry, cortical thickness measurement, basal forebrain volumetry, and multi-variate statistics have emerged [7,8]. The availability of large image data repositories to the neuroimaging community has necessitated the development of distributed data and processing infrastructures to access data and online image analysis tools and to assess longitudinal brain changes [9–12].

Many efforts have been directed at creating brain image repositories including the recent US Alzheimer Disease Neuroimaging Initiative (ADNI) [13]. Numerous efforts, such as NeuroLOG [14]and Neurogrid [15], have been conducted which focus on providing grid infrastructures that support neuroimaging application [16]. At present, however, these applications tend to be either focused on specific pathologies or are directed at supporting a subset of neuroimaging applications. Moreover, these solutions are tightly bound to specific platforms, which may limit their wider adoption across neuroscience. neuGRID is an effort which targets the limitations of existing neuroimaging based Grid infrastructures and aims to provide an infrastructure and a set of complementary

* Corresponding author. Tel.: +44 013 325 91881.
E-mail address: ashiq.anjum@cern.ch (A. Anjum).

0925-2312/$ - see front matter
http://dx.doi.org/10.1016/j.neucom.2013.01.042

analysis services that are designed to support and enhance research. neuGRID is an EC-funded effort which will allow the collection and archiving of large amounts of imaging data paired with services, Grid-based algorithms and computational resources. The major benefit will be faster discovery of new disease markers that will be valuable for earlier diagnosis and development of innovative drugs.

It needs to be stressed that some of the presently available algorithms can take many hours per brain to run on a state-of-the-art workstation [17]. The modus operandi today is that of scientists physically migrating image data to remote imaging centres where they can find expertise and computational facilities for analysing small personal datasets (a few hundreds of images at most). Typically, a research fellow can spend months at an image analysis centre where he/she learns the use of the algorithms on personal image data, then returns to the original research group, where he/she can install all or part of the procedure and run jobs either in house or remotely on the image analysis centre servers. This scenario is becoming unsustainable and it needs to change radically in the near future. Conventional file sharing mechanisms e.g. peer-to-peer file sharing, can be used to share image and clinical data, however such mechanisms still require the researchers to feed in the data to computational analysis programmes. The benefits of such data sharing on a Grid based infrastructure include the fact that the data remains *online*, it can be shared across organisational boundaries through the concept of virtual organisations in the Grid, better resource utilisation through Grid scheduling and better access control.

Neuroimaging researchers require infrastructures that can enable the large-scale computation of standardised pipelines on large datasets provided by the major data repositories. Domain researchers also require an infrastructure that enables collaborative studies that may involve multiple geographically dispersed research centres. However efficiently optimising the neuroimaging pipelines that are both compute and data intensive on an e-Science infrastructure poses various challenges. First, these pipelines consist of a large number of tasks. The CIVET pipeline [18], for instance, can consist of 108 tasks and the workflow turn-around time is around 8 h for a single brain scan. Secondly, these pipelines can generate a large amount of data. CIVET has been shown to produce ten times more data than it consumes [19]. This can add up to several terabytes for larger studies and several months of computations. Thirdly, neuroimaging pipelines consist of a large number of fine-grained tasks that have shown to severely affect the turn-around time of the workflow. However workflow optimisation methods have not kept pace with the rise of complexity in workflows, hence researchers have called for new approaches to optimising, managing and enacting them. Moreover, they need multi-criteria optimisation methods that can effectively optimise workflows for computation.

To achieve a low turn-around time (compute optimisation), computations within a workflow must be distributed in order to benefit from parallelism. On the other hand, to achieve data efficiency computations must be localised in order to limit expensive network transfers. We used a multi-objective meta-heuristic to optimise scientific workflows and evaluated through a number of real world scientific workflows – focusing on the CIVET [18] workflow in particular.

The domain of multi-objective meta-heuristics has been an active area of research [20] and various successful applications have been reported. For instance, several multi-objective evolutionary approaches have been used to optimise distributed computing capabilities such as scheduling [21] and classification [22]. However their use in the optimisation of scientific workflows has not been explored. Since the compute and data performance may be dependant on various factors, the search space of all possible optimised workflow plans may be large. An evolutionary meta-heuristic, being a stochastic population-based search algorithm, enables the simultaneous exploration of a search space as members of a population can be randomly distributed across the search space. Moreover, the genetic operations of mutation and crossover can enable the fine-grained control of the balance between exploitation (the ability to leverage characteristics of known solutions) and exploration (the ability to explore new parts of the search space). Multi-objective evolutionary algorithms (MOEAs) regarded as state-of-the-art include the Non-dominated Sorting Genetic Algorithms II (NSGA-II) [23], Strength Pareto Evolutionary Algorithm 2 (SPEA2) [24], Indicator based Evolutionary Algorithm (IBEA) [25] and HyPE [26].

In this paper we present work on the set of intelligent services in the neuGRID project that has been specified in consultation with its user community and developed to facilitate neuroimaging analysis, such as Alzheimer's studies. The services, using machine learning approaches, can intelligently plan, execute and 'glue' a spectrum of user applications to a range of available Grid platforms thereby creating a foundation for pervasive cross-platform services for neuroimaging analysis and promoting interoperability between diverse projects in this domain. This paper provides the background for understanding the characteristics of scientific analyses, highlights the issues that influence their optimisation and presents an approach for their intelligent planning and execution.

2. A service oriented analysis environment in neuGRID

In order to facilitate analysis and collaboration that can address the community's requirements, a service oriented analysis environment has been proposed in neuGRID in which high-level distributed services such as querying, workflow management, provenance, and anonymisation services [27] coordinate and interact to support user analyses. Such services will help the users in sharing data and knowledge and should enrich medical decision support systems [28]. The preferred approach is to implement a service oriented architecture (SOA) [29]. The service layer in neuGRID was implemented using the SOA paradigm in order to have a flexible and reusable medical services layer, which can be customised for various applications. The following paragraphs illustrate how the services in the neuGRID analysis environment, will coordinate to facilitate the neuroimaging analysis process, using SOA principles. In the later sections we emphasise how these services can intelligently support the analyses and improve the planning and execution process on distributed infrastructures.

The first action in the neuroimaging analysis cycle, shown in Fig. 1, is to register images in the neuGRID store that have been collected from the hospital data acquisition systems or have been imported from other research projects. The border in Fig. 1 denotes the limit of the neuGRID project infrastructure as determined from users' requirements. As an example, consider a new clinical site that may wish to make use of the neuGRID infrastructure to share data within a wider research community. Existing data would be put through a process that enforces quality control, anonymisation and ethical compliance. The data is then integrated with the neuGRID data model, which enables other researchers to access it and carry out their research. As new datasets are acquired they go through a local quality control step before passing through the same system-wide quality control, formatting, ethical compliance and data model integration processes.

Once the data has been registered, the next step in the analysis process is to make the data browsable through automated querying tools. Consequently, an appropriate data access mechanism needs to be put in place. For example, a researcher may be interested in a rare form of a disease and may want to carry out a statistically significant analysis. However, the researcher's