GeoComputation 2009

Yong Xue[1,2], Forrest M. Hoffman[3], and Dingsheng Liu[4]

[1] State Key Laboratory of Remote Sensing Science, Jointly Sponsored by the Institute of Remote Sensing Applications of Chinese Academy of Sciences and Beijing Normal University, Institute of Remote Sensing Applications, Chinese Academy of Sciences,
P. O. Box 9718, Beijing 100101, China

[2] Department of Computing, London Metropolitan University, 166-220 Holloway Road,
London N7 8DB, UK

[3] Oak Ridge National Laboratory, Computational Earth Sciences Group, Building 5600,
Room C221, MS 6016, P.O. Box 2008, Oak Ridge TN 37831-6016, USA

[4] Center for Earth Observation and Digital Earth, Chinese Academy of Sciences,
No.45, Bei San Huan Xi Road, Beijing, China

y.xue@londonmet.ac.uk, forrest@climatemodeling.org,
dsliu@ceode.ac.cn

1 Preface

The tremendous computing requirements of today's algorithms and the high costs of high-performance supercomputers drive us to share computing resources. The emerging computational Grid technologies are expected to make feasible the creation of a computational environment handling many PetaBytes of distributed data, tens of thousands of heterogeneous computing resources, and thousands of simultaneous users from multiple research institutions (Giovanni *et al.* 2003).

The Workshop on GeoComputation continues with the ICCS conferences held in Amsterdam (2002), St. Petersburg (2003), Krakow (2004), Atlanta (2005), Reading (2006), Beijing (2007) and Krakow (2008). GeoComputation is about using various different types of geographical and environmental data and developing relevant tools within the overall context of a computational scientific approach. It is concerned with new computational techniques, algorithms, and paradigms that are dependent upon and can take advantage of Grid Computing. It includes spatial data analysis, dynamic modeling, simulation, space-time dynamics and visualization and virtual reality. This conference will offer presentations from a variety of sources, both local, national and international and will enable you to network with others working in similar fields.

Grid computing technology is a new method for processing remotely sensed data. Jianwen Ai et al. in their paper “Grid Workflow Modeling for Remote Sensing Retrieval Service with Tight Coupling” discusses some application cases based on Grid computing for Geo-sciences and the application limit of Grid in remote sensing, and provides a method for Grid Workflow modeling for remote sensing. Tight-coupling remote sensing algorithms cannot be scheduled by a Grid platform directly. Therefore, we need an interactive graphical tool to present the executing relationships of algorithms and to generate automatically the corresponding submitted description files for a Grid platform.

Image resampling, which is frequently used in remote sensing processing procedures, is a time-consuming task. Parallel computing is an effective way to speed up

G. Allen et al. (Eds.): ICCS 2009, Part II, LNCS 5545, pp. 345–348, 2009.

this processing; however, recent parallel image resampling algorithms with massive time-consuming global processes like I/O, always lead to low efficiency and non-linear speedup ratios, especially when the number of computing nodes increases beyond a certain extent. And what's more, the various geo-referencing related to different processing applications cause a real problem for code reuse. To solve these problems, PIRA-PIO (Parallel Image Resampling Algorithm with Parallel I/O) algorithm, an asynchronous parallelized image resampling algorithm with parallel I/O, is proposed in the paper by Ma *et al.* "An Asynchronous Parallelized and Scalable Image Resampling Algorithm with Parallel I/O". Parallel I/O on parallel file systems and asynchronous parallelization using I/O hidden policy to sufficiently overlap the computing time with I/O time is used in PIRA-PIO for performance enhancement. In addition, the design of reusable code like design pattern will be used for improving flexibility in different remote sensing image processing applications. Through experimental and comparative analysis, its outstanding parallel efficiency and perfect linear speedup is shown in this paper.

Jingshan Li *et al.* in their paper "Design and Implementation of a Scalable General High Performance Remote Sensing Satellite Ground Processing System on Performance and Function" discuss design and implementation of a scalable high performance remote sensing satellite ground processing system using a variety of advanced hardware and software application technology for performance and function. These advanced technologies include the network, parallel file system, parallel programming, job scheduling, workflow management, design patterns, etc., which make the performance and function of remote sensing satellite ground processing systems scalable enough to fully meet the high performance processing requirements of multi-satellite, multi-tasking, massive remote sensing satellite data. The "Beijing-1" satellite remote sensing ground processing system is introduced as an instance.

Remote sensing data plays a key role in understanding complex geographic phenomena. Clustering is a useful tool in discovering interesting patterns and structures within multivariate geospatial data. One of the key issues in clustering is the specification of an appropriate number of clusters, which is not obvious in many practical situations. In this paper "Incremental Clustering Algorithm for Earth Science Data Mining" Ranga Raju Vatsavai provided an extension of a G-means algorithm which automatically learns the number of clusters present in the data and avoids over estimation of the number of clusters. Experimental evaluation on simulated and remotely sensed image data shows the effectiveness of their algorithm.

The booming of Earth observation provides decision-makers with more available geospatial data as well as more puzzles about how to understand, evaluate, search, process, and utilize those overwhelming resources. The paper from Wang *et al.* "Overcoming Geoinformatic Knowledge Fence: An exploratory of intelligent geospatial data preparation within spatial analysis" introduce a concept termed geoinformatic knowledge fence (GeoKF) to discuss the knowledge-aspect of such puzzles and an approach to overcoming them. Based on analysis of the gap between common geography sense and geoinformatic professional knowledge, the approach comprises analysis of space modeling and spatial reasoning to match decision models to the online geospatial data sources they need. Such approaches enable automatic and intelligent searching of suitable geospatial data resources and calculating their suitability to a given spatial decision and analysis. An experiment with geo-services, geo-ontology and rule-based reasoning

(Jess) is developed to illustrate the feasibility of the approach in scenarios of data preparation within decisions of bird flu control.

Service Oriented Architecture is not widely used in GIS. Although there are a few organizations that have launched Service Oriented GIS applications, it is not possible for all interested parties to utilize those services. Because those are proprietary organizations, users have to pay large amounts of money for those services. Sometimes they do not always provide the desired services to the users either. Therefore users have to move to another application. It is really a waste of money and time. The paper "Service Oriented Customizable Framework to Manipulate GIS Data" from Ranasinghe and Karunaratne aims to provide a solution by designing a service oriented customizable framework to manipulate GIS data.

Spatial relations play an important role in computer vision, scene analysis, geographic information systems (GIS) and content-based image retrieval. Fuzzy Allen relations are used to define the fuzzy topological relations between different objects and to detect object positions in images. In the paper "Spatial Relations Analysis by Using Fuzzy Operators" from Salamat and Zahzah, fuzzy aggregation operators are used for information integration along with polygonal approximation of objects. This new approach offers low temporal and computational complexity for the extraction of topological and directional relations.

A wide variety of data mining techniques are being applied to the growing body of Earth Science data. From small scale measurement data to global climate simulation output, very large or long time series databases of environmental and climate data are proving difficult to analyze and interpret. Data mining techniques--like cluster analysis, principle components analysis (PCA), classification and regression tree (CART) analysis, and neural networks--are being applied to problems of feature extraction, model-data comparison, and validation/verification. However, the size and complexity of Earth Science data are stretching the limits of commercial statistical packages and many freely available analysis tools, which were not designed to scale up to terabyte- and petabyte-sized datasets. Scalable statistical tools are needed to run on very large parallel supercomputers in order to analyze data of this size. The following papers address these issues by demonstrating how data mining techniques can be applied in the Earth Sciences and by describing innovative computer science techniques and methods that can support analysis and discovery in Earth Sciences.

Increasingly large datasets acquired by NASA for global climate studies demand larger computation memory and higher CPU speed to extract useful and revealing information. While boosting the CPU frequency is getting harder, clustering multiple lower performance computers thus becomes increasingly popular. This prompts a trend of parallelizing the existing algorithms and methods by mathematicians and computer scientists. In the paper "A Parallel Nonnegative Tensor Factorization Algorithm for Mining Global Climate Data", Zhang *et al.* take on the task of parallelizing the Nonnegative Tensor Factorization (NTF) method, with the purpose of distributing large datasets across cluster nodes, thus reducing the demand on a single node, blocking and localizing the computation at the maximal degree, and finally minimizing the memory use for storing matrices or tensors by exploiting their structural relationships. Numerical experiments were performed on a NASA global sea surface temperature dataset and resulting factors are analyzed and discussed.

The ultimate goal of data visualization is to clearly portray features relevant to the problem being studied. This goal can be realized only if users can effectively communicate to the visualization software important features of interest. To this end, Johnson *et al.* describe in the paper "Querying for Feature Extraction and Visualization in Climate Modeling" two query languages used by scientists to locate and visually emphasize relevant data in both space and time. These languages offer descriptive feedback and interactive refinement of query parameters, which are essential in any framework supporting queries of arbitrary complexity. They apply these languages to extract features of interest from climate model results and describe how they support rapid feature extraction from large datasets.

The recurrence of periodic environmental states is important to many systems of study, and particularly to the life cycles of plants and animals. Periodicity in parameters that are important to life, such as precipitation, are important to understanding environmental impacts, and changes to their intensity and duration can have far reaching impacts. To keep pace with the rapid expansion of earth science datasets, efficient data mining techniques are required. Discrete Fourier transform (DFT) and wavelet analysis are useful data mining tools for rapidly searching for changes in the intensity of seasonal, annual, or interannual events by projecting the magnitude and shift of periodicities onto power spectrum plots. Brooks explores the strengths and limitations of DFT and wavelet spectral analysis using output from the Parallel Climate Model (PCM). Spectral analysis is used to diagnose model behavior, and locate land surface cells that show shifting cycle intensity, which could be used as an indicator of climate change. Example routines in Octave/Matlab and IDL are provided.

Danek's paper "Seismic wave field modeling with graphics processing units" describes the GPGPU - general-purpose computing on graphics processing units, which is a very effective and inexpensive way of dealing with time consuming computations. In some cases even a low end GPU can be a dozens of times faster than a set of modern CPUs. Utilization of GPGPU technology can make a typical desktop computer powerful enough to perform necessary computations in a fast, effective and inexpensive way. Seismic wave field modeling is one of the problems of this kind. Sometimes one modeled common shot-point gather or one wave field snapshot can reveal the nature of an analyzed wave phenomenon. On the other hand these kinds of models are often a part of complex and extremely time consuming methods with almost unlimited needs for computational resources. This is always a problem for academic centers, especially now when times of generous support from oil and gas companies have ended.

Acknowledgment

We would like to thank our reviewers, Stefano Furin (Università di Ferrara, Dipartimento di Scienze della Terra), William W. Hargrove (U.S. Department of Agriculture – Forest Service), Jian Huang (University of Tennessee), Phillip Kegelmeyer (Sandia National Laboratories), Vipin Kumar (University of Minnesota), G.Q. Li (CAS, China), Kamesh Madduri (Lawrence Berkeley National Laboratory), Richard T. Mills (Oak Ridge National Laboratory), George Ostouchov (Oak Ridge National Laboratory), Christopher T. Symons (Oak Ridge National Laboratory), Raju R. Vatsavai (Oak Ridge National Laboratory), and Xingquan "Hill" Zhu (Florida Atlantic University), for sharing their kind support and expertise to improve the content of this workshop.