Vis Geosci (2006) 10: 1–12
DOI 10.1007/s10069-005-0023-z

ORIGINAL ARTICLE

Dave A.Yuen · Benjamin J. Kadlec · Evan F. Bollig
Witold Dzwinel · Zachary A. Garbow
Cesar R. S. da Silva

Clustering and visualization of earthquake data in a grid environment

Received: 24 August 2004 / Revised: 15 February 2005 / Accepted: 16 February 2005 / Published online: 21 May 2005

Abstract We present a web client-server service WEB-IS, which we have developed for remote analysis and visualization of seismic data consisting of both small magnitude events and large earthquakes. We show that the problem-solving environment (PSE) intended for prediction of large magnitude earthquakes can be based on this WEB-IS idea. The clustering schemes, feature generation, feature extraction techniques and rendering algorithms form a computational framework of this environment. On the other hand, easy and fast access both to the seismic data distributed among distant computing resources and to computational and visualization resources can be realized in a GRID framework. We discuss the usefulness of NaradaBrokering (iNtegrated Asynchronous Real-time Adaptive Distributed Architecture) as a middleware, allowing for flexibility and high throughput for remote visualization of geophysical data. The WEB-IS functionality was tested both on synthetic and the actual earthquake catalogs. We consider the application of similar methodology for tsunami alerts.

Keywords Earthquakes · Seismic data · Visualization · Clustering · PSE · Grid computing

D. A.Yuen · B. J. Kadlec (✉) · E. F. Bollig
Department of Geology and Geophysics,
University of Minnesota,
Minneapolis, MN 55455-219 USA
E-mail: Kadlec@msi.umn.edu

D. A.Yuen · B. J. Kadlec (✉) · E. F. Bollig · Z. A. Garbow
C. R. S. da Silva
Minnesota Supercomputing Institute,
University of Minnesota, Minneapolis,
MN 55455–0219 USA

W. Dzwinel
AGH Institute of Computer Science,
al. Mickiewicza 30, 30–059 Kraków, Poland

C. R. S. da Silva
Faculdade de Computacão,
Universidade de Santo Amaro,
Rua Isabel Schmidt, 349 São Paulo,
Brazil

Introduction

Earthquake dynamics is one of the most challenging problems in modern geophysics and computer modeling. In the last fifty years a great deal of effort has been devoted trying to devise techniques for their prediction. Unfortunately, in many cases these methods are unreliable or work only in retrospective studies. This has created some skepticism in the geophysical community about the predictability of earthquakes (Song and Simons 2003).

A predictive knowledge can be acquired either by constructing models expressed in terms of the partial differential equations, finite automata, or supervised learning techniques such as feed-forward neural networks. For earthquakes and other SOC (self-organized criticality) processes these approaches may not be adequate as a precise predictive tool. Instead, the knowledge can be extracted from past seismic measurements by using methods of pattern recognition (Jain and Dubes 1988; Ertoz et al. 2003) and then used for extrapolating the future. The process of knowledge extraction consists of studying similarity and dissimilarity of N-dimensional (N-D) feature vectors of measurements. These vectors build-up spatio-temporal clusters made of similar objects (Dzwinel et al. 2003; Dzwinel et al. 2005). We can expect that these clusters, their shapes and mutual position in the N-D feature space, encode the knowledge about the seismic patterns in the region under interest. Of course, multi-resolution cluster structures depend strongly on many environmental properties and on background noise. The success of pattern recognition requires elimination of all of these factors and detection of clusters closely related to the earthquake precursors.

The proper selection of both feature generation and clustering techniques are critical for knowledge extrac-

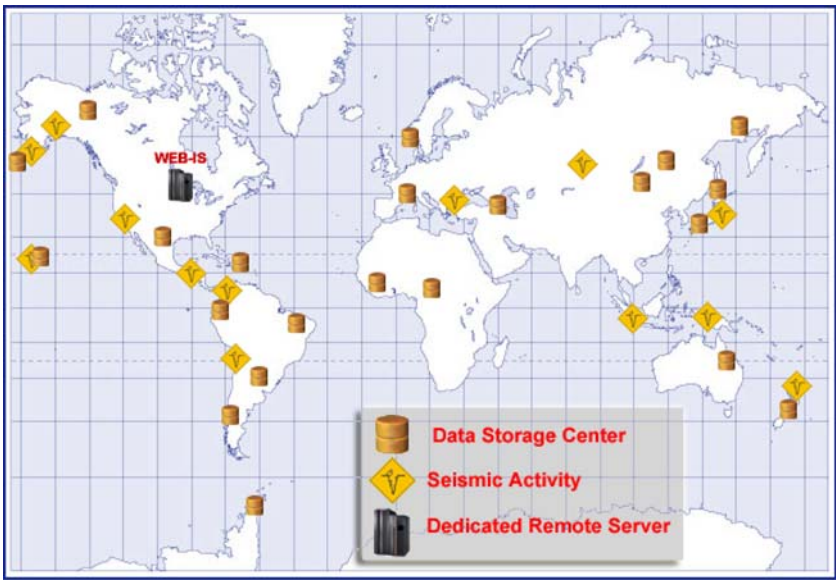

Fig. 1 Map showing distribution of data acquisition centers at regions of high seismic activities, supercomputing centers where this data is stored and analyzed, and WEB-IS as a solution to integrate all of these locations

tion. In Dzwinel et al. (2003 and 2005) we show that different techniques of clustering in different types of feature spaces have to be combined to scrutinize various aspects of the earthquake dynamics. Therefore, integrated pattern recognition software similar to CLUTO (Rasmussen et al. 2003; Karypis 2003), must be used for exploration of earthquake data catalogs. Moreover, to explore the relevance of the clusters obtained, the clustering package must be empowered with advanced visualization tool such as Amira (2004). This visualization package should allow for:

1. On-line exploration of the feature space.
2. Selection of the most appropriate clustering scheme, clustering parameters and a proper similarity measure.
3. Representation of data in a compressed, compact and understandable format.
4. Results on-demand for quick analysis.

This idea of predicting earthquakes by observation of similarities between thousands and millions of seismic events by visualization of earthquake clusters requires the fast access to large databases. The largest earthquake catalogs comprise GBytes of data. Taking into account also the data from tsunami earthquakes and micro-earthquakes in mines, the total amount of data collected by seismic centers spread all over the world is huge. Moreover, the knowledge extraction of earthquake precursors may demand exploration of cross-correlations between different catalogs. Consequently, both fast communication between data centers and large disk spaces are required.

As shown in Fig. 1, earthquake modeling centers, data mining centers and data acquisition centers, which collect earthquake data from regions with high seismic activity, are spread all over the world. Therefore, the unprocessed data needs to be stored and then transferred to a dedicated remote server for data processing. After processing, the results should be returned to the data acquisition centers and/or other clients. Broad access to remote pattern recognition and visualization facilities allows for scrutinizing local data catalogs by using peer-to-peer connections of data acquisition centers to data preprocessing servers. Clients in the network can compare various types of earthquake catalogs, data measured in distant geological regions, and the results from various theoretical and computational models. By comparing data accessible in the network we have a chance to eliminate environmental factors and to extract the net earthquake precursory effects.

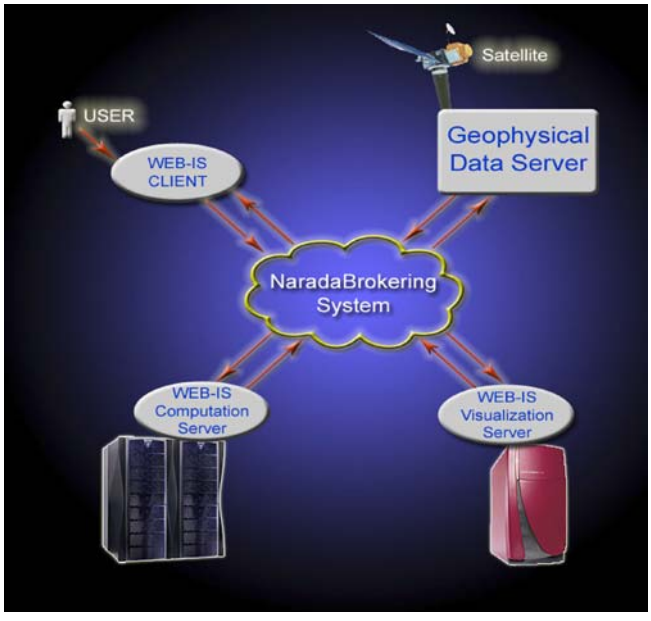

Fig. 2 The NaradaBrokering middleware allows the WEB-IS client to communicate with multiple servers, in the form of a GRID service, for data visualization and analysis

Integration of a variety of hardware, operating systems, and their proper configuration results in many communication problems between data centers. Efficient, reliable, and secure integration of distributed data and software resources, such as pattern recognition and visualization packages, is possible only within the GRID paradigm of computing.

The GRID mode of computing has flourished rapidly in recent years and has facilitated collaboration and accessibility to many types of resources, such as large data sets, visualization servers and computing engines. Scientific teams have developed easy-to-use, flexible, generic and modular middleware, enabling today's applications to make innovative use of global computing resources. Remote access tools were also produced to visualize huge datasets and monitor performance and data analysis properties, effectively steering the data processing procedures interactively.

We discuss the idea of an integrated problem-solving environment (PSE) intended for the analysis of seismic data and prediction of the earthquakes. A simplified scheme of the data acquisition and visualization system for geophysical data, integrated within a NaradaBrokering environment is displayed in Fig. 2. This system promotes portability, dynamic results on-demand, and collaboration among researchers separated by long distances by using a client server paradigm. To accomplish this we provide a light-weight front-end interface for users to run locally while the a remote server takes care of intensive processing tasks on large data bases, off-screen rendering and data visualization. We present a web service WEB-IS, which we have developed for remote visualization of clustered seismic data.

In the first section we present typical earthquake data we use in our system. The earthquake clustering and visualization idea are discussed in the following section. We then describe the WEB-IS functionality and the usefulness of NaradaBrokering (iNtegrated Asynchronous Real-time Adaptive Distributed Architecture) as a middleware allowing for better flexibility and higher throughput for such a GRID service. In the conclusions we summarize our work and show the perspectives of the idea presented.

Table 1 Observed (JMA) and synthetic (AUMF) earthquake data specifications

Catalog	Total number of events	Time interval (years)	Events with $m > 6$	Earthquake magnitude
JMA data	42,370	5	62	$3 < m < 7.90$
U	32,185	150	32	$3.26 < m < 6.68$
A	25,881	150	30	$3.26 < m < 6.73$
F	10,475	150	16	$3.43 < m < 6.73$
M	29,039	150	20	$3.41 < m < 6.81$

Earthquake data

There exist various types of seismic data catalogs: local and global, synthetic and observed. There is a definite need to reveal their complexity, eliminate noise [e.g. using wavelets (Erlebacher and Yuen 2004)], outliers and extract the knowledge hidden behind the patterns the data create.

There are two main sources of earthquake data: computer modeling and seismic measurements, which produce synthetic and actual data catalogs, respectively. The earthquake catalogs can comprise GBytes of data when including data from various kinds of earthquakes such as regular, tsunami (and tsunamigenis) and micro-earthquakes in mines.

The usefulness of synthetic catalogs based on theoretical models, such as those listed in Table 1, depends on the extent to which the models mimic realistic fault activity including earthquakes. The use of synthetic data for analysis of seismic events and developing methodology for prediction of earthquakes has many advantages, all of which are exceedingly relevant for revealing complex physical behavior. Moreover, synthetic data have the advantage of retaining the statistical reliability of the results. The data are free of measurement errors, which occur in estimating earthquake magnitudes and hypocentral locations, and do not suffer from incomplete recording of small events, which exist in natural catalogs. Synthetic data generated by computational models can comprise many events covering large spatial areas and extremely long time spans. These features are very attractive both for studying new methods for data analysis and understanding theoretical aspects of earthquake dynamics. As an example of synthetic data we use the results from the numerical model of a discrete fault plane simulated with 3D elastic dislocation theory and power-law creep for the central San Andreas Fault (Eneva and Ben-Zion 1997a and 1997b; Ben-Zion 1996). The results cover the earthquake distribution in space, time and level of magnitude. We study catalogs from four different models representing various levels of fault zone disorder. These are models with statistically uniform brittle properties (U), with a Parkfield type Asperity (A), with fractal brittle properties (F), and with multi-size-heterogeneities (M). These models and various statistical properties of the catalogues have been discussed in greater detail elsewhere (Eneva and Ben-Zion 1997a and 1997b; Ben-Zion 1996). The time interval covers every event, which occurred during the last 150 years of simulated fault activity and this period contains from 1 to 3×10^4 events.

The data representing seismic activities of the Japanese islands collected by the Japan Meteorological Agency (JMA) (Ito and Yoshioka 2002) is an example of measured data. The seismic events were detected during a 7-year time interval from 1997 to 2003. The data set consists of 4×10^4 seismic events with magnitudes m, position in space (latitude X, longitude Y, depth z) and

occurrence time t. The lowest magnitudes were determined by using a detection level, estimated from the Gutenberg-Richter frequency-size distribution. Thus the statistical completeness of the earthquakes above the detection level assures that there is no significant lack of events in space and time.

Every seismic event i can be represented as the data point $\mathbf{x_i}$ in N-dimensional space, i.e., $\mathbf{x}_i \in \mathbf{R}^N$ and $\mathbf{x}_i = \{x_{i1}, x_{i2}, ..., x_{iN}\}$ where x_{il} are the features describing the earthquake. These features can be represented by earthquake characteristic values: energy, moments, position or other geophysical parameters resulting from theoretical or/and statistical models. The clustering idea allows for grouping similar (or correlated) objects and finding their common properties. The clustering and vector quantization empowered with visualization facilities studying seismic patterns, and their uniformity, replication, or periodicity. Visual inspection greatly assists in detecting subtle structures, which escape the classical pattern extraction algorithms such as clustering techniques. In the following section we discuss the idea of earthquake clustering and visualization.

Clustering of earthquakes and visualization

Cluster analysis (Gowda and Krishna 1978; Jain and Dubes 1988) — unsupervised learning — divides data objects, or feature vectors, into groups (clusters), for the purposes of summarization, improved understanding and knowledge extraction.

Definition

Let X will be a data set: $X = \{\mathbf{x}_i\}_{i=1,...,M}$ and where $\mathbf{x} \in \mathbf{R}^N$ and ; $\mathbf{x}_i = \{x_{i1}, x_{i2}, ..., x_{iN}\}$. We define as an *m-clustering* of X, as the partition of X into m clusters $C_1, ..., C_m$ provided three conditions are met:

- $C_i \neq 0$, i = 1,...,m
- $U_{i=1,...,m} C_i = X$
- $C_i \cap C_j = 0$, $i \neq j$, j = 1,...,m

The basic that steps must be followed in order to develop a clustering task are the following:

1. *Feature selection* – Features must be properly selected to encode as much information as possible. Parsimony and minimum redundancy among the features is a major goal.
2. *Proximity measure* – The measure of how "similar" (or "dissimilar") two features vectors are.
3. *Clustering criterion* – This defines what constitutes a good clustering structure in the given data set depends on the interpretation of the term "sensible", depending on the type of clusters are expected in the data set e.g., such as oblate, elongated, "bridged", circular etc.
4. *Clustering algorithms* – Choose a specific algorithmic scheme that unravels the clustering structure of the data set.
5. *Validation and interpretation of results* – The final processes of clustering involving the knowledge of the specialist who analyzes the results.

There are two principal types of clustering algorithms: non-hierarchical and agglomerative schemes (Gowda and Krishna, 1978). Because the global information about the data is required for the non-hierarchical schemes, they non-hierarchical algorithms suffer from a high computational complexity and most of them require an a priori knowledge about of the number of clusters.

Agglomerative clustering schemes consist in the subsequent merging of smaller clusters into the larger clusters, basing on proximity and clustering criteria. Depending on the definition of these criteria, there exist many agglomerative schemes such as: average link, complete link, centroid, median, minimum variance and nearest neighbor algorithm. The hierarchical schemes are very fast for extracting localized clusters with non-spherical shapes. The proper choice of proximity and clustering criteria depend on many aspects such as dimensionality of data.

For example, by using smart clustering criterion in molecular dynamics (da Silva et al. 2002) we can provide a more efficient method for clustering N-dimensional data (where $N < 5$) in situations with extremely high computational complexity and spatial locality between data events. Two feature vectors are taken as linked if they satisfy some proximity measure criterion and are spatially closer to each other than some maximum cut-off radius. The method performs the cluster calculation in two steps. In the first step it uses a data structure named Verlet's Neighbor List (VNL) (Theodoris and Koutroumbas 1998). In order to avoid quadratic time complexity for the creation of the VNL, a Linked Cell (LC) structure algorithm is used (Theodoris and Koutroumbas 1998). Creating an LC and subsequently a VNL from the LC is very fast, with a time complexity of order N (da Silva et al. 2002). The feature vectors and the VNL form a compact representation of a graph. Consequently, the second step is simply a backtracking search for connected trees in that graph. In this step, we have the opportunity to reject the inclusion of some feature vector in a given tree if the vector does not satisfy some additional criterion. This method can be applied in higher dimensions, thus reducing the number of feature vector components by using principal component analysis (PCA) or non-linear multidimensional scaling (Ertoz et al. 2003).

Additionally, we note that all agglomerative algorithms suffer from the lack of properly defined control parameters, which can be matched for the data of

interest and hence can be regarded as invariants for other similar data structures.

There are four basic directions in which clustering is of use (Gowda and Krishna 1978):

1. *Data reduction* — Data (or features) can be quantized into groups represented by only one representative feature vector instead of many similar ones belonging to a specific group (cluster).
2. *Hypothesis generation* – Applying cluster analysis to a data set in order to infer some hypotheses concerning the nature of the data. Thus, clustering can be used as a vehicle to suggest hypotheses.
3. *Hypothesis testing* – Cluster analysis can be used for verification of a specific hypothesis' validity, on the basis of common properties of feature vectors belonging to a specific cluster.
4. *Prediction based on groups* – If an anonymous feature vector belongs to a specific cluster characterized by a set of common attributes, this vector will be also described by these attributes.

In (Dzwinel et al. 2003; Dzwinel et al. 2005) we used clustering for all four purposes. We investigated multi-resolution earthquake patterns in both observed data from Japanese islands and the synthetic data generated by numerical simulations. At the highest resolution, we analyzed the local cluster structures in the 4-D data space of seismic events. We demonstrated that small magnitude events produce local spatio-temporal patches that correspond to large neighboring events. Seismic events, quantized in space and time, generate the 7-D feature space characterized by the earthquake parameters. Using a non-hierarchical clustering algorithm and multi-dimensional scaling, we explore the multitudinous earthquakes by real-time 3-D visualization and inspection of the multivariate clusters. At the spatial resolutions characteristic of earthquake parameters, all ongoing seismicity, both before and after the largest events, accumulates to a global structure consisting of a few separate clusters in the feature space. We show that by combining the clustering results from low and high resolution spaces, we can recognize precursory events more precisely and unravel vital information that cannot be discerned at a single resolution. The problem of finding clusters in data is challenging, when multidimensional clusters are of widely differing sizes, densities and shapes, and when the data contains large amount of noise and outliers. Therefore, we used clustering schemes, which are based on nearest neighbor proximity measures, such as mutual nearest neighbor agglomerative clustering (Theodoris and Koutroumbas 1998) and modern shared nearest neighbor schemes (Ertoz et al. 2003).

As shown in Fig. 3 the software environment of a Problem Solving Environment (PSE) for clustering and visualization of seismic data can be constructed on the basis of multi-dimensional clustering software and Multi-Dimensional Scaling (MDS) procedures (Jain and Dubes 1988; Theodoris and Koutroumbas 1998) for visualization. Our goal is to construct an interactive system for data-mining, allowing users to match the most appropriate clustering schemes on the structure of actual seismic data. The proposed PSE combines local clustering techniques in the data space with a non-hierarchical clustering in the feature space. The raw data are represented by M, n-dimensional vectors $\mathbf{f}_i$ of measurements f_{ik} $(i=1...M,\ k=1...n)$ (such as coordinates, depth, magnitude etc.). The data space can be searched for patterns and visualized by using local or remote pattern recognition with advanced visualization capabilities. The data space $\mathbf{X}=\{\mathbf{f}_i\}$ is transformed to a new abstract space $\mathbf{F}$ of vectors $\mathbf{F}_j$ $(j=1...M)$. The coordinates F_{jl} $(l=1...N)$ of these vectors represent nonlinear functions of measurements F_{jl}, which are averaged in given space-time windows. This transformation allows for quantization and amplification of data and their characteristic features, neglecting both the noise and other random components. The new features F_{jl} form N-dimensional feature space. We use MDS procedure (Dzwinel, 1994) for visualizing the multi-dimensional events in 3-D space. This transformation allows for a visual inspection of the N-dimensional feature space. The visual analysis helps greatly in detecting subtle cluster structures not recognized by classical clustering techniques, and selecting the best pattern detection procedure used for data clustering by classifying the anonymous data and formulating new hypotheses.

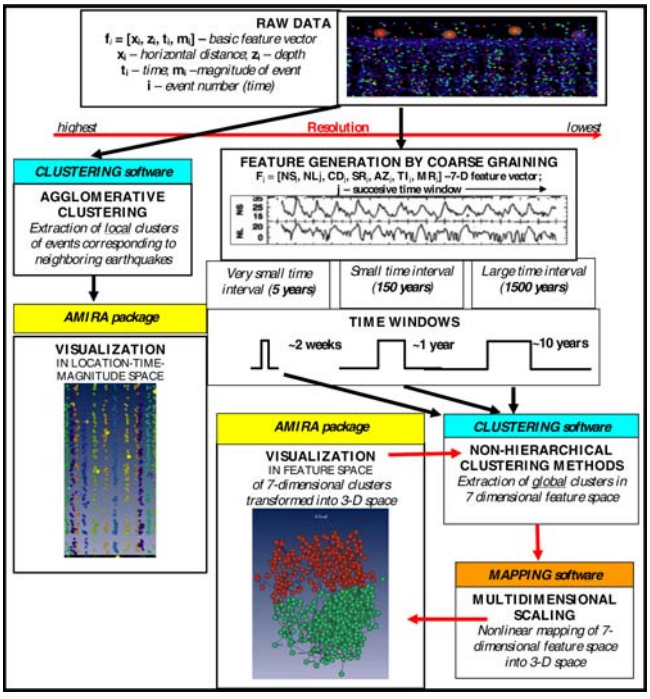

Fig. 3 The diagram of multi-resolution analysis of seismic events

In the visualization part of our PSE we employ different methods for visualizing earthquake clusters depending on the type of data. Earthquake events taken

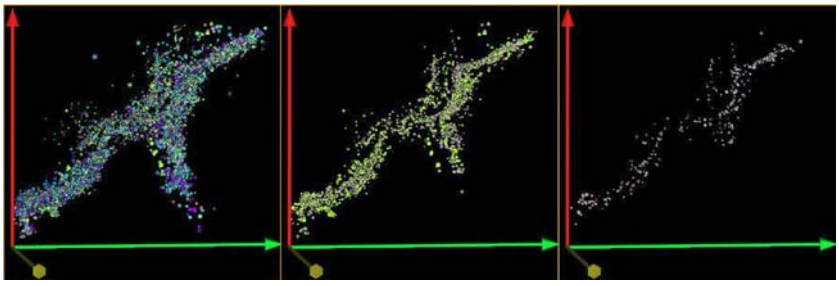

Fig. 4 JMA data from the island of Japan shown at three different time scales. This allows for observation of time slice dependencies occurring in the clustered data. From left to right: 6 years, 1 year, 1 month
Movie 1: This movie shows time dependant clusters based on the JMA data from Japan

from our synthetic data sets are plotted as spheres in 3-dimensional space; two dimensions of position and one dimension of time account for the coordinates of this visualization space. Applying color-mapping and radius scaling to these spheres allows cluster-code and magnitude classifications to be accounted for as well. In this representation, the user can navigate through the clusters by zooming, rotating, and translating the clustered data in its 3-dimensional space. Plotting time as a third dimension allows dependencies between events to be explored by flying through this static spatio-temporal environment.

When 3-dimensions of position are available for visualization, as with the real earthquake data collected by JMA, a dynamic view needs to be used to incorporate time into the visualization. One method we use for creating this type of dynamic visualization is showing clusters at different time-scales according to defined units of time (year, month, day, etc.). Choosing these different tracks of time is like viewing the data through a filter, allowing time-dependencies to be explored relative to the scale. We employ another method for showing how clusters form in continuous subsets of time. By stepping through the entire range of time at specified intervals, we can see how these clusters are formed. This is like viewing a slideshow and watching clusters appear as earthquakes bond with their neighbors. (Fig. 4)

Having the idea of earthquake data analysis and visualization we can implement it as an integrated computer system. However, first we have to decide how to adapt it as the Problem Solving Environment accessible by clients spread all over the world (see Fig. 1).

NaradaBrokering in GRID computing

In general, large datasets and high-performance computing resources are distributed across the world. When collaboration and sharing of resources are required, a computational GRID infrastructure needs to be in place to connect these servers. There must exist protocols available to allow clients to tap into these resources and harness their power. Computational grid can be seen as a distributed system of ''clients'', which consists of either ''users'' or ''resources'' and proxies. A GRID can be implemented using an event brokering system designed to run on a large network of brokering nodes. Individually, these brokering nodes are competent servers, but when connected to the brokering system, they are able to share the weight of client requests in a powerful and efficient manner. Examples of this include GRID Resource Brokering (Ferreira et al. 2002) and NaradaBrokering.

NaradaBrokering (NB) (2004) is a distributed event brokering system designed to run on a large network of cooperating broker nodes. As shown in Fig. 5, the framework of this system is hierarchical where a broker (server) is part of a cluster of brokers, which is part of super-cluster, and the framework continues expanding until all brokers are part of one universal cluster in a tree-like structure. Clusters consist of tightly connected brokers with multiple links to brokers from other clusters. This allows for selecting alternate communication routes during failures. This type of organizational scheme is called a ''small world network'' or ''scale-free network'' (Barabási and Bonabeau 2003) where the average communication path length between brokers increases logarithmically while the network size increases geometrically. In uncontrolled situations the respective growth has exponential character. The ''scale-free network'' architecture allows NB to support large heterogeneous client configurations that can scale to arbitrary size.

This type of GRID architecture suits well to the WEB-IS functionality as a PSE for earthquake data analysis and as an integrated computational environment for data exchange and common ventures. The seismic data centers from the networking point of view represent complex hierarchical cluster structure. They are located geographically in the regions of the high seismic activity within heavily populated areas of economic importance. Therefore, the seismic data centers create distant superclusters of various ''density'' of computational resources corresponding to the size and importance of the regions. These superclusters are sparse in the sense of computational resources devoted for earthquake detection and data acquisition. However, these same structures contain important computational, scientific and visualization facilities with strong interest in the analysis of earthquake data and earthquake modeling. The efficient interconnection of these sites is of principal interest. Due to the ''small world network'' structure of NB it is possible to select the most efficient routing schemes, considerably short-

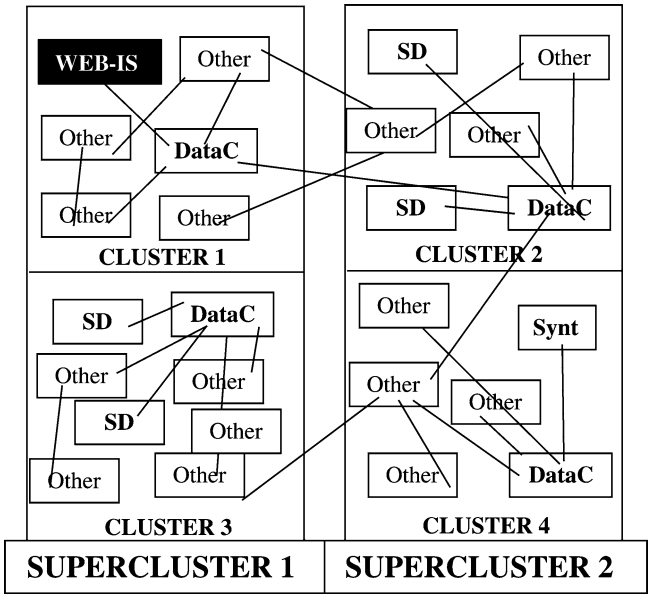

Fig. 5 WEB-IS as a client in a NaradaBrokering network. SD (seismic data), DataC (data servers), Synt (synthetic data). SD and Synt publish the data, which are subscribed by WEB-IS server and other clients. WEB-IS publishes the results of data analysis and visualization that are retrieved by subscribed clients SD and Synt. For simplification we show a Two-Supercluster Structure. In a realistic seismic network the communicated clients can be spread over many levels of superclusters. However, due to the "small world network" structures of NaradaBrokering, the average communication path length between brokers increases only logarithmically

ening the average communication path length between brokers.

Clients need to communicate with NB through a Java Messaging Service (Fox and Pallickara 2002) application, which is based on the publish/subscribe model. In this model, clients send requests to a distributed network of brokers, who then determine the most capable server to complete the request. This messaging system of NB is appropriate to link the clients, both users and resources, together.

Construction of efficient and user friendly Problem Solving Environment requires integration of data analysis and visualization software with NB environment, in such a way that it can be easily accessed via the Internet. We created an integrated data interrogation toolkit, which we call WEB-IS (Wang et al. 2004).

The WEB-IS (Integrated System) as a problem solving environment

This WEB-IS toolkit provides three modules which each specialize in distinct forms of visualization and data analysis. We focus here on the first two modules, WEB-IS1 and WEB-IS2. WEB-IS1 is a software tool that allows remote, interactive visualization and analysis of large-scale 3-D data over the Internet (Yuen et al. 2004) through the interaction between client and server. The second module WEB-IS2, which is also based on the client-server paradigm, utilizes the immense visualization capabilities of the software package Amira (2004) to provide remote analysis of various types of data (Wang et al. 2003). WEB-IS acts as a PSE through a web portal used to solve problems by visualizing and analyzing geophysical datasets, without requiring a full understanding of the underlying details in software, hardware and communication (Garbow et al. 2002; Kadlec et al. 2003).

As shown in Fig. 6, the primary goal of WEB-IS in the geosciences is to provide middleware that sits between the modeling, data analysis tools and the display systems that local or remote users access. In the case of large and physically distributed datasets, it is necessary to perform some preprocessing and then transmit a subset of the data to one or more processes or visualization servers to display. The details of where and how the data migrates should be transparent to the user. WEB-IS makes available to the end users the capability of interactively exploring their data, even though they may not have the necessary resources such as sufficient software, hardware or datasets at their local sites. This method of visualization allows users to navigate through their rendered 3-D data and analyze for statistics or apply data mining techniques (Grossman et al. 2001), such as cluster analysis (Dzwinel et al., 2003). Also, utilization of the powerful 3-D visualization package Amira, provides the ability to remotely analyze, render, and view large datasets, such as 3-D mantle convection and seismic tomography models over the Internet (Wang 2003).

In WEB-IS, multiple clients can harness the power of a large visualization server to supply 3-D image rendering and data analysis capabilities. Figure 7 shows the client–server software infrastructure of WEB-IS (Integrated System) and how the system allows the client to access data on the server. To the client, the process of accessing and manipulating the data appears simple and robust, while the middleware takes care of the network communication, security and data preparation.

A mixture of programming languages went into the design of WEB-IS1 to achieve optimal performance, and to provide the highest quality middleware service. The server utilizes C/C++ for the fast performance, Mesa3D library for off-screen rendering of OpenGL worlds, and Python scripts to interact with the server machine. On the client side, WEB-IS appears as a Java Applet, which is embedded into a simple HTML web page.

The WEB-IS1 Client provides the front end Graphical User Interface (GUI) for our homegrown earthquake clustering and visualization server. As depicted in Fig. 8, the Java applet GUI maintains platform independency (Garbow et al. 2003). Using icons and mouse controls, rather than text entry at a command line, the applet lets users productively interact with the system. A remote computer visualizes the data off-screen, while the client-side applet merely displays the result. Users can manipulate the data by clicking and dragging their

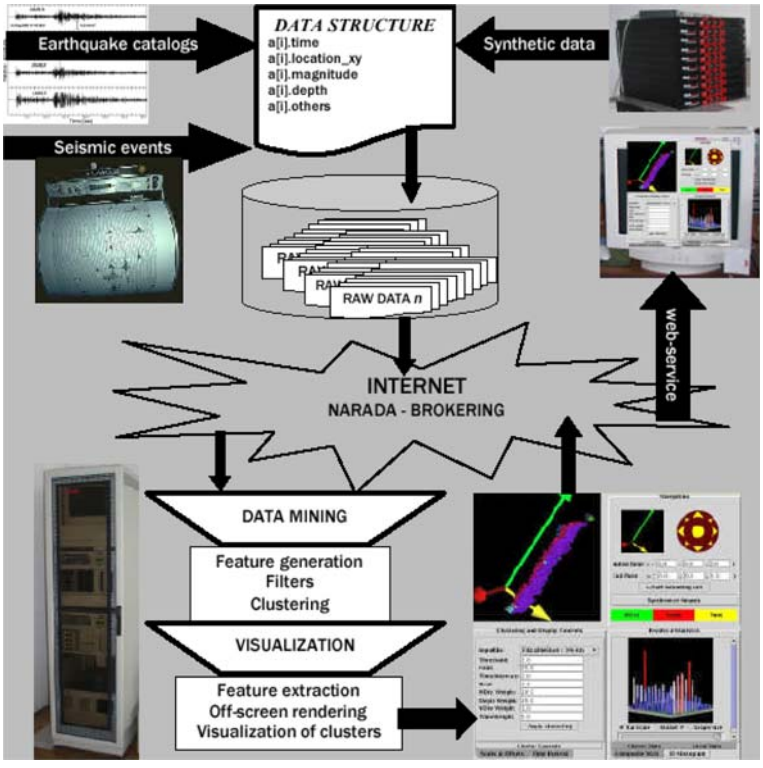

Fig. 6 The conceptual scheme of WEB-IS functionality

mouse over the Server Image. The Navigation Panel works like a remote control to the Server Image for creating a selection box and for gathering user input on rotating, zooming, and translating the image. In addition, the Clustering & Display Controls allow users to manipulate clustering parameters, adjust the axis scales, and broaden or narrow the time interval with specified parameters. The Results and Statistics Panel show both numerical statistics and a 3-D histogram of the clustered dataset. As a side note, the 3-D histogram, as well as the remote control axis, take advantage of jGL (2004), a Java based OpenGL package, for rendering local 3-D worlds.

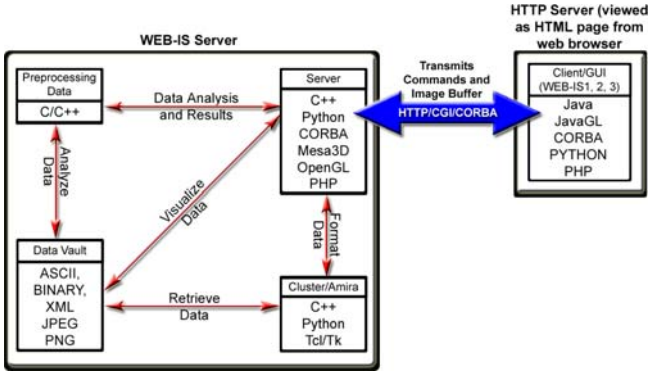

Fig. 7 WEB-IS Software Scheme: The user interacts with the Java applet (right), which provides the client-side GUI and a front-end for all interactions. The server (left) receives requests from the client, and performs these processor intensive tasks, returning the results to the client to be displayed

While WEB-IS2 shares Amira's ability to handle many different file types [see (Amira 2004) for more details], we chose to follow a simple data format in WEB-IS1 to keep everything as simple as possible. Our first requirement is data must be contained in an ASCII file containing the ''.dat'' suffix. The second specification is data must be singular events (i.e. earthquake events). The first line of the data file is considered to be a ''header'', containing a single integer value declaring the file's dimensions (at this point only a value of 4 or 5 will work). The event data begins on the second line in the format X, Y, Z, Time (optional), and Magnitude. The event properties must be readable as float (float or int values work), and either tabs or spaces must separate them. As in most visualization toolkits, X, Y and Z indicate the physical axis (3-D world) with which the property is associated. The Time and Magnitude properties are self-explanatory and are visualized as a visible/invisible event depending on the time frame and the event ball size respectively. It is important to note that for 2-D datasets which contain Time and Magnitude (i.e. synthetic data), we chose to automatically visualize Time on the Z-axis assuming users' bothered by this approach could reformat their data to contain a constant value in the Z column of the dataset.

WEB-IS2 acts as an interactive application that allows users on a client machine to control, display, and share visual information generated by a remote server running Amira. WEB-IS2 saves users from creating their own customized visualization programs by harnessing the visualization power of prewritten software. The server in this client-server scheme is a standard version

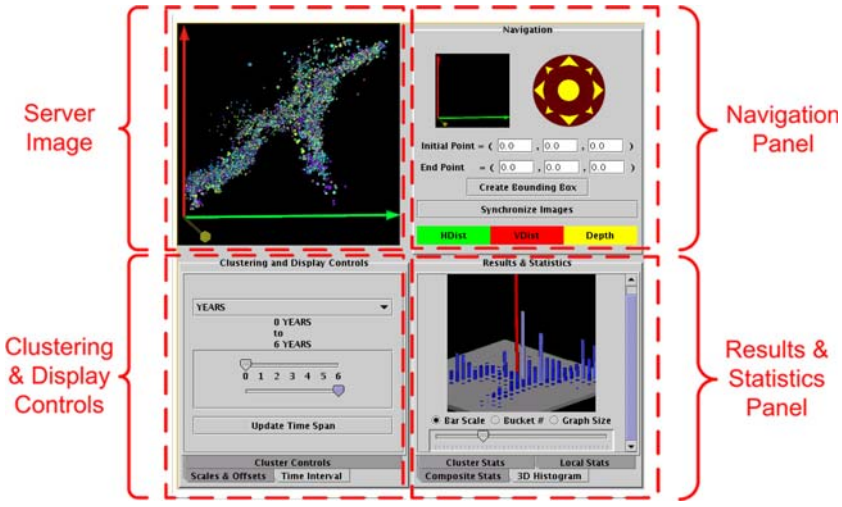

Fig. 8 WEB-IS1 (http://boy.msi.umn.edu/web-is1) allows users to navigate through their rendered 3-D data and interactively analyze the data for statistics or apply data mining techniques, such as cluster analysis

of Amira, manipulated over the Internet by a Java applet. Most of the computational work is performed on the server, allowing the client to concentrate on user interface instead of intensive data processing. From the perspective of visualization, Amira has already proven to be a very effective tool. Users construct programs visually with flow charts of modules that represent tasks or via Tcl (Tool Command Language 2004) scripts. In both cases, the result is the display of a subset of the original data (see Fig. 9 below). Furthermore, Amira uses some advanced algorithms that capitalize on the hardware available on many of the current commodity graphics cards. Amira has proved to be valuable in visualizing geophysical phenomena such as numerical simulations of 3D convection in the Earth's mantle (Dubuffet et al. 2000; Dubuffet et al. 2002).

The current version of WEB-IS1 (developed by us from the ground up) implements the SOAP protocol in a move to create a true web-service providing decentralized clients and servers. SOAP, or Simple Object Access Protocol, is a relatively lightweight, XML based protocol that has no concept of a central server, making it ideal for message handling in a distributed computing environment (Englander 2002; Snell et al. 2002). Several toolkits and APIs currently exist to make development of SOAP/XML services easier, including three utilized by WEB-IS1: gSOAP (2003), Apache SOAP (2004) and Apache Axis (2004). gSOAP, created at Florida State University, was designed to optimally bind C/C++ and SOAP, which makes it ideal for server-side use. Apache SOAP, developed by the Apache SOAP community, is an implementation of the SOAP v1.1 and SOAP Messages with Attachments specifications for Java. The follow-on project to Apache SOAP, named Apache Axis, is essentially a framework for constructing SOAP processors such as clients and servers. WEB-IS1 uses the Apache Axis API to easily create a SOAP handling client that composes messages based on the Apache SOAP API. Because of the standardization of the SOAP protocol, messages are easily passed across platforms and languages. This has allowed us to retain our multi-language web-service and preserve the basic structure and features found in the original WEB-IS1 system. Unfortunately, retaining state using SOAP is very difficult for most uses. Since only one server may run per port on a machine, there is a good possibility that multiple clients will share a single server process. This can potentially be seen as both a problem and a benefit depending on the application. In the geosciences, this sharing allows users to easily collaborate across vast distances (see Fig. 1).

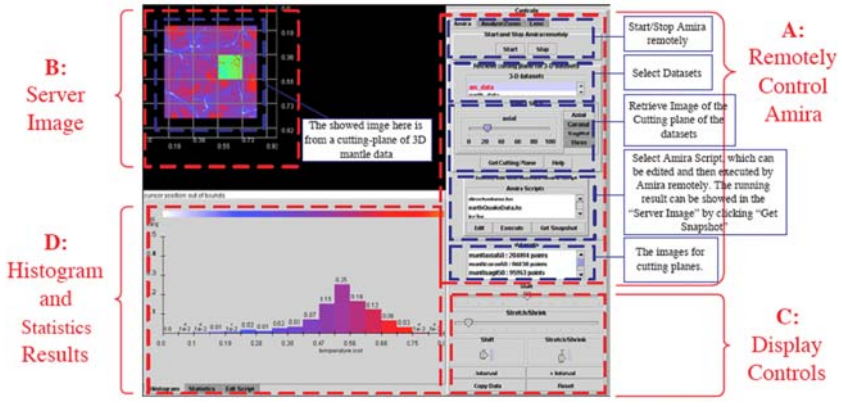

Fig. 9 Example of WEB-IS2 in action

Any clustering or transformation changes done to the visualization are immediately available to all users connected to the same port, and because the client is a Java applet, users can collaborate from any java capable web-browsers. Unfortunately, everyone shares the same data and visualization; this prevents users from clustering and visualizing data without interference.

The conversion of our system to handle SOAP messaging allowed for easy integration of the NaradaBrokering (NB) message-oriented middleware (MoM). A Java proxy designed for passing messages into NB sits between the client applet and the NaradaBrokering system. This proxy listens for messages from the server and immediately provides the client with any messages published to a requested topic. Another proxy waits on the server-side for any messages from a client published on NB, sending the messages to the port on which the server is connected. On the client side a Java proxy listens to both the WEB-IS1 client and NB, and passes the messages in the correct direction. Figure 10 shows a screenshot of the WEB-IS1 system using NB. It is important to note that none of the components need to run on the same machine since they all communicate through ports and sockets. The proxies are also not needed if the client or server is written to send messages directly to NB. We use them here for simplicity in message passing.

Testing of the NaradaBrokering system has shown that multiple clients, servers, and/or proxies receive the same message when subscribed to a common topic. This property is convenient, but one must beware that NaradaBrokering simply passes messages across machines, and it is up to the subscriber to handle them correctly. If a proxy is designed to pass all messages through to a client, and the client is not prepared to receive a message, the proxy must have the ability to keep messages queued or dispose of them after a timeout period. Also, since every subscriber receives the same message, there is a possibility of distinctly different servers receiving a common message, producing different results, and then publishing those results to the same topic as a response. In a situation such as this, a client will receive the result produced by the first responding server, yet the result may not coincide with its request.

Security in Web/GRID Services is a hot topic today, receiving a large amount of attention from the research world. We recognized early that there was a potential security threat posed by transferring data to a WEB-IS server for analysis. Since a WEB-IS server must handle remote datasets, some form of data upload must be completed before analysis can begin. However, due to the requirement for the WEB-IS server to be installed in the root web script directory (cgi-bin), there is a possibility that a malicious script could be uploaded causing damage to both hardware and software on the server. Unfortunately, unless WEB-IS is deployed on every data server in the world, this threat cannot be prevented. We also realized that many server administrators would prefer to minimize the number of possible entry points for data on their machines. Therefore, we chose to prohibit data upload via the WEB-IS system, depending instead on other common secure transfer protocols typically found on all systems [i.e. Secure FTP (sftp) and Secure Copy (scp)]. This essentially minimizes the threat of an attack through the WEB-IS system, and it installs a trust-based security where administrators grant explicit permission for users to upload data. Since this also limits the functionality of WEB-IS as a global resource for all researchers to use on demand, we provide the WEB-IS source code for users to install on their local machines. The source, available on the homepage (http://webis.msi.umn.edu/), also leaves open the possibility for users to implement their own form of data upload.

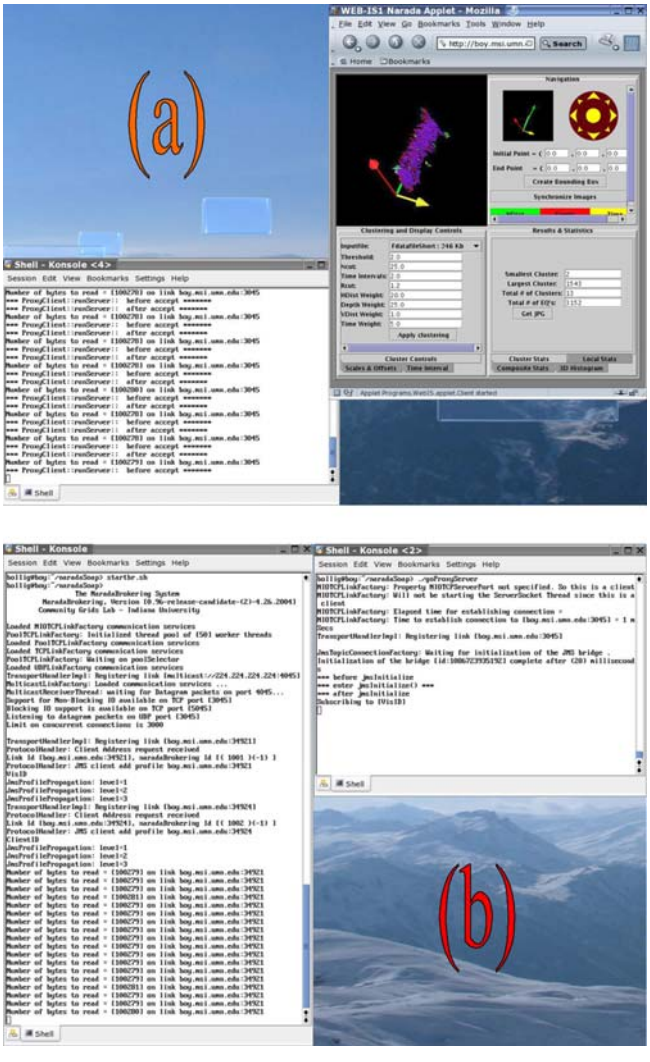

Fig. 10a, b Here we see screenshots of the WEB-IS1 system using NaradaBrokering from both the client-side **(a)** and sever-side **(b)**

Conclusions and perspectives

In this paper we presented the philosophy and primary engines of our system:

1. Data analysis tools such as pattern recognition software and clustering techniques;
2. High performance visualization packages;
3. The NaradaBrokering grid environment;
4. WEB-IS integration toolkit.

We demonstrated that these components are not simply vague ideas or prototypes in the very beginning stage of testing. They already exist; they can work both independently and coupled in a single special purpose system. This system can be developed creating the backbone of the sophisticated computational data acquisition environment, which can be devised for specific needs of the geophysical community. Integration of measurement devices and geophysicists must be the principal goal for the nearest future. Equipped with only PDAs or laptops, and working on location in unreachable desert terrains with remote data acquisition centers or perhaps just analyzing data in one of the many computation facilities located around the globe, geophysicists will be enabled unlimited access to data resources spread all over the world. Specialized computation and visualization facilities will empower geophysicists with robust virtual equipment allowing not only for fast in situ communication, but also for data visualization, and comparing data obtained in other locations through the use of sophisticated pattern recognition software. The reverse interaction of this system, i.e., its application for data mining and realization of large geophysical projects must not be underestimated.

We see the principal goal of our work in contributing to the construction of a global warning system, which can be used for prediction of catastrophes such as various types of earthquakes along the circum Pacific belt, where there is a great concentration of people. For example, similar methodology can be used for tsunami earthquake alerting. Theoretical models of faulting and seismic wave propagation, used for the computation of radiated seismic energy from broadband records at teleseismic distances (Boatwright and Choy 1986) can be adapted to the real-time situation when neither the depth nor the focal geometry of the source is known accurately. The distance-dependent approximation was used in Newman and Okal (1998). By analyzing some singular geophysical parameters (such as the energy-to moment ratio Θ [Newman and Okal 1998)] for regular earthquakes, the results obtained from the theoretical models agree well with values computed from available source parameters (e.g., as published by the National Earthquake Information Center). It appears however that the so called ''tsunami earthquakes'' - characterized by the significant deficiency of moment release at high frequencies—yield the values of Θ considerably different (greater on more than 1 unit) than those obtained for the regular earthquakes. Thus Θ value can be used as a suitable criterion for discriminating various types of earthquakes in a short duration of time, like an hour. However, this hypothesis holds only for a few cases. For, so called, ''tsunamigenic earthquakes'' this difference is not so clear. Moreover, the value of the moment computed on the base of long-period seismic waves can be underestimated. For example, analysis of the longest period normal modes of the Earth, ${}_0S_2$ and ${}_0S_3$, excited by the December 26, 2004 Sumatra earthquake (Stein and Okal 2004), yields an earthquake moment of 1.3×10^{30} dyn-cm, approximately three times larger than the 4×10^{29} dyn-cm measured from long-period surface waves. Therefore, instead of a single-value discrimination we recommend using more parameters (dimensions) for detecting tsunami earthquakes. As shown in Okal et al. (2003) and Walker et al. (1992), one could employ other T-phase characteristics such as its duration, seismic moment, and spectral strength or even similar features associated with the S-phase.

We believe that the lack of success in predicting earthquakes still comes from the lack of communications between researchers and difficulties in free and fast access to the various types of data. Therefore, we hope that globalization of computation, data acquisition and visualization resources, together with fast access through a scale-free network facilitated by systems such as NaradaBrokering, will provide a triumphant solution to this problem.

Acknowledgements We would like to thank Yunsong Wang for contributing ideas and figures, Zhenyu Lu for his help with NaradaBrokering, and Dr. Gordon Erlebacher for always presenting us with new ideas. We thank Professors Seth Stein and Emile Okal for discussions concerning the excitation of the normal modes in the Sumatran earthquake. WD acknowledges support from the Polish Committee for Scientific Research (KBN) Grant No. 3T11C05926. This research has been supported by Math-Geo and ITR programs of the National Science Foundation and the Digital Technology Center of the Univ. of Minnesota.

References

Amira (2004) — Advanced 3D Visualization and Volume Modeling, http://www.amiravis.com

Apache Axis (2004) http://ws.apache.org/axis/

Apache SOAP (2004) http://ws.apache.org/soap/

Barabási AL, Bonabeau E (2003) Scale-Free Networks. Scientific American, Vol. 288 pp 60–69

Ben-Zion Y (1996) Stress, slip, and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations. J Geophys Res 101(B3):5,677–5,706

Boatwright J, Choy G L (1986) Teleseismic estimates of the energy radiated by shallow earthquakes, J Geophys Res 91:2,095–2,112

da Silva CRS, Justo JF, Fazzio A (2002) Structural order and clustering in annealed a-SiC and a-SiC:H, Physical Review B, Volume 65, 104108

Dubuffet F, Rabinowicz M, Monnereau M (2000) Multiple scales in mantle convection, Earth and Planetary Science Letters 178(3–4):351–366

Dubuffet F, Yuen DA, Rainey ESG (2003) Controlling thermal chaos in the mantle by positive feedback from radiative thermal

conductivity, Nonlinear Processes in Geophysics 9:311–323, 2002
Dzwinel W (1994) How to Make Sammon's Mapping Useful for Multidimensional Data Structures Analysis, Pattern Recognition 27(7):949–959
Dzwinel W, Yuen DA, Kaneko YJBD, Boryczko K, Ben-Zion Y (2003) Multi-resolution clustering analysis and 3-D visualization of multitudinous synthetic earthquakes, Visual Geosciences 8:12–25 http://link.springer.de/link/service/journals/10069/contents/tfirst.htm
Dzwinel W, Yuen DA, Boryczko K, Ben-Zion Y, Yoshioka S, Ito T (2005) Nonlinear multidimensional scaling and visualization of earthquake clusters over space, time and feature space, Nonlinear Processes in Geophysics 12:117–128
Eneva M, Ben-Zion Y (1997a) Techniques and parameters to analyze seismicity patterns associated with large earthquakes, Journal of Geophysical Research 102/B8:785–795
Eneva M, Ben-Zion Y (1997b) Application of pattern recognition techniques to earthquake catalogs generated by model of segmented fault systems in three-dimensional elastic solids. Journal of Geophysical Research 102/B11:513–528
Englander R (2002) Java and SOAP, O'Reilly & Associates, Inc., Sebastopol, CA
Erlebacher G, Yuen DA (2004) A wavelet toolkit for visualization and analysis of large datasets in earthquake research. Pure and Applied Geophysics 161:1–15
Ertoz L, Steinbach M, Kumar V (2003) Finding Clusters of Different Size, Shapes and Densities in Noisy, High-Dimensional Data, Army High Performance Center, technical report, April
Ferreira L et al. (2002) Introduction to Grid Computing with Globus, IBM Redbook series, IBM Corporation, http://ibm.-com/redbooks
Fox G, Pallickara S (2002) JMS Compliance in the Narada Event Brokering System, International Conference on Internet Computing, pp. 391–402
Garbow ZA, Erlebacher G, Yuen DA, Sevre EO, Nagle AR, Kaneko Y (2002) Web-Based Interrogation of Large-Scale Geophysical Datasets and Clustering Analysis of Many Earthquake Events from Desktop and Handheld Devices, American Geophysical Union Fall Meeting Abstract
Garbow ZA, Yuen DA, Erlebacher G, Bollig EF, Kadlec BJ (2003) Remote Visualization and Cluster Analysis of 3-D Geophysical Data over the Internet Using Off-Screen Rendering, to appear in Visual Geosciences
Gowda CK, Krishna G (1978) Agglomerative clustering using the concept of nearest neighborhood, Pattern Recogn 10:105
Grossman RL, Kamath C, Kegelmeyer P, Kumar V, Namburu RR (Eds.) (2001) Data Mining for Scientific and Engineering Applications, Kluwer Academic Publish, 605 pp
gSOAP (2003): C/C++ Web Services and Clients, http://www.cs.fsu.edu/~engelen/soap.html
Ito T, Yoshioka S (2002) A dike intrusion model in and around Miyakejima, Niijima and Kozushima. Tectonophysics 359:171–187
Jain D, Dubes RC (1988) Algorithms for Clustering Data. Prentice-Hall Advanced Reference Series
jGL (2004) — 3D Graphics Library for Java, http://nis-lab.is.s.u-tokyo.ac.jp/~robin/jGL/
Kadlec BJ, Yang XL, Wang Y, Bollig EF, Garbow ZA, Yuen DA, Erlebacher G (2003) WEB-IS (Integrated System): An Overall View, Eos. Trans. AGU, 84(46), Fall Meet. Suppl., Abstract NG11A-0163
Karypis G (2003) Cluto: Software Package for Clustering High-Dimensional Datasets, http://www-users.cs.umn.edu/~karypis/cluto/, November
The NaradaBrokering (2004) Project @ Indiana University, http://www.naradabrokering.org
Newman AV, Okal EA (1998) Teleseismic estimates of radiated seismic energy: the S/M_0 discriminant for tsunami earthquakes, J Geophys Res 103/B11:23,885–23,898
Okal EA, Alasset P-J, Hyvernaud O, Schindelé F (2003) The deficient T waves of tsunami earthquakes, Geophys J Int 152:416–432
Rasmussen MD, Deshpande MS, Karypis G, Johnson J, Crow JA, Retzel EF (2003) "wCLUTO: A Web-Enabled Clustering Toolkit", Plant Physiology Vol 133, No 2, pp 511 October
Snell J, Tidwell D, Kulchenko P (2002) Programming Web Services with SOAP. O'Reilly & Associates, Inc., Sebastopol, CA: January
Song TA, Simons M (2003) Large trench-parallel gravity variations predict seismogenic behavior in subduction zones, Science, 301:630–633
Stein S, Okal E (2004) Ultra-long period seismic moment of the great December 26, 2004 Sumatra earthquake and implications for the slip process, http://www.earth.nwu.edu/people/emile/research/Sumatra.pdf
Theodoris S, Koutroumbas K (1998) Pattern Recognition, Academic Press, San Diego
Tool Command Language (2004), http://www.tcl.tk
Walker DA, Mc Creery CS, Hiyoshi Y (1992) T-phase spectra, seismic moment and tsunamigenesis, Bull Seism Soc Am 82:1,275–1,305
Wang Y (2003) Visualization Web Service, Master of Science Thesis, Florida State University, Fall
Wang Y, Erlebacher G, Garbow ZA, Yuen DA (2003) Web-based service of a visualization package "Amira" for the geosciences, Visual Geosciences
Wang Y, Bollig EF, Kadlec BJ, Garbow ZA, Erlebacher G, Yuen DA, Rudolph M, Yang LX, Sevre EOD (2004) WEB-IS (Integrated System): An Overall View, Submitted to Visual Geosciences
Yuen DA, Garbow ZA, Erlebacher G (2004) Remote data analysis, Visualization and Problem Solving Environment (PSE) Based on Wavelet Analysis in the Geosciences, Visual Geosciences 8:83–92 DOI10.1007/x10069–003–0012-z