A Seismic Data Management and Mining System

Sotiris Brakatsoulas and Yannis Theodoridis

Computer Technology Institute,
P.O. Box 1122, GR-26110 Patras, Greece
http://www.cti.gr/RD3/DKE

Abstract. A Seismic Data Management System should meet certain requirements implied by the nature of seismic data. This kind of data is not solely characterized by alphanumeric attributes but also from a spatial and a temporal dimension (the epicenter and the time of earthquake realization, respectively). Moreover, visualizing areas of interest, monitoring seismicity, finding hidden regularities, and assisting to the understanding of regional historic seismic profiles are essential capabilities of such a system. Thus, a spatiotemporal DBMS, a user-friendly visualization system, and a set of data analysis and knowledge discovery techniques compose, to our opinion, a so-called Seismic Data Management and Mining System (SDMMS). In this paper, we describe SDMMS requirements and present a prototype that, further to the above, aims to integrate seismic data repositories available over the WWW.

1 Introduction

Seismic data are phenomena recorded by seismologists in order to describe and study tectonic activity. For a certain time period, tectonic activity can be described by recording geographic information, i.e. epicenter and disaster areas, together with attributes like magnitude, depth of epicenter, etc. Desirable components of a seismic data application include tools for quick and easy data exploration and inspection, techniques for generating historic profiles for specific geographic areas and time periods, techniques providing the association of seismic data with other geophysical parameters of interest such as soil profile, geographic and perhaps specialized maps (e.g. topological and climatological) for the presentation of data to the user and, topline, visualization components supporting sophisticated user interaction.

In particular, we distinguish three user profiles that such an application should support:

- researchers of geophysical sciences, who could, for example, be interested in constructing seismic profiles of certain areas and time periods or in discovering regions of similar seismic behavior.
- key personnel in public administration, usually needing information like distances between epicenters and and other geographical entities like schools and heavy industries.
- web surfers, who may query the system for seismic properties of general interest, e.g. for finding all epicenters of earthquakes in distance no more than 50Km from Athens.

Management of seismic data, due to their spatiotemporal nature, demands more than a relational DBMS and a GIS on top of it. Recent advances in the areas of Databases, Knowledge Discovery in Databases (KDD) and Data Visualization allow better approaches both for their efficient storage/retrieval and analysis. Commercial DBMS's have already started providing tools for the management of spatiotemporal data, while KDD techniques have shown their potential through a wide range of applications. Additionally, several research prototypes have applied these technologies and have utilized certain benefits.

Our proposal consists not only of a spatiotemporal DBMS, but also of a data mining module that incorporates methods for classification, for finding clusters and for finding association rules. This module aims to the better understanding of seismic phenomena and of the relationships between seismic parameters themselves or between them and other factors like subsoil properties, weather conditions during earthquake realization etc. It could be a useful tool for geophysical scientists who are more interested in high level concepts than in collections of raw observational data.

Moreover, it is our intention to face the challenge of exploiting the available seismic data repositories over the web. Assuming that the user would periodically, but not very frequently, check these repositories for newly available data, load them into the local database (namely, in a batch update fashion) and taking into consideration the very large size and heterogeneity of these repositories, it seems natural to adopt a Data Warehouse (DW) approach for the implementation of our prototype.

The rest of the paper is organized as follows. In the next section we survey available technologies and research trends that we intend to include in the SDMMS prototype. Section 3 presents the SDMMS prototype under development and describes its architecture, functionality and current status of implementation. In section 4, we discuss related work and present several research prototypes developed for management of spatiotemporal and geophysical data. Finally, section 5 concludes with directions for future work.

2 Requirements for Management and Mining of seismic data

A combination of three state-of-the-art database technologies is required for the efficient handling of seismic data, namely, spatiotemporal databases, data warehouse and data mining techniques.

2.1 Spatiotemporal Databases

Modelling the real world for seismic data applications requires the use of spatiotemporal concepts like snapshots, changes of objects and maps, motion and phenomena [14]. In particular, we are concerned with the following concepts:

- *Spatial objects in time points.* It is a simple spatiotemporal concept where we record spatial objects in time points, or, in other words, we take snapshots of them. This concept is used, for example, when we are dealing with records including position (latitude and longitude of earthquake epicenter) and time of earthquake realization together with attributes like magnitude, depth of epicenter, and so on.

- *Spatial objects in time intervals.* This could be the case when we intend to capture the evolution of spatial objects over time, for example when, additionally to the attributes mentioned previously, we are interested in recording the duration of an earthquake and how certain parameters of the phenomenon vary throughout the time interval of its duration.
- *Layers in time points.* Layers correspond to thematic maps showing the spatial distribution of certain attributes in the database. The combination of layers and time points results into snapshots of a layer. For example, this kind of modelling is used when we are interested in magnitude thematic maps of earthquakes realized during a specific day inside a specific area.
- *Layers in time intervals.* This is the most complex spatiotemporal concept we are interested in for the modelling of phenomena. For example, modelling the whole sequence of earthquakes, including the smaller in magnitude that precede or follow the main earthquake, uses the notion of layers in time intervals.

It is clear that we are mostly interested in the spatiotemporal attributes of earthquake data. For example, typical queries that involve the spatial and the temporal dimension of data are the following:

- *Find ten epicenters of earthquakes realized during the past four months and reside more closely to a given location.*
- *Find all epicenters of earthquakes residing in a certain region, with a magnitude* $M > 5$ *and a realization time in the past four months.*

In order to support the above data models and queries, new data types [4] and access methods have been proposed in the literature. Seismic data are multi-dimensional and, as a consequence, require different techniques for their efficient storage and retrieval than those traditionally used for alphanumeric information. A great deal of effort has been spent for the development of efficient spatial access methods (SAMs) and some (R-trees, Quad-trees) have been integrated into commercial DBMS (Oracle, Informix, DB2). Recently, SAMs that take the dimension of time into consideration have been also proposed [16, 12, 13]

2.2 Data warehouses

Additional to the spatiotemporal DBMS, for the reasons stated in section 1, a data warehouse approach can be adopted for the integration of the available sources of seismic data over the Web and the utilization of on-line analytical processing (OLAP) technology. A data warehouse is usually defined as a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decision making process [5]. We illustrate the benefits obtained by such an approach with two examples of operations supported by spatial data warehouse and OLAP technologies:

- A user may ask to view part of the historic seismic profile, i.e. the ten most destructive quakes in the past twenty years, over Europe, and, moreover, can easily view the same information over Greece (more detailed view, formally a *drill-down* operation) or over the whole world (more summarized view, formally a *roll-up* operation).

– Given the existence of multiple thematic maps, perhaps one for quake magnitude and one for another, non-geophysical parameter such as the resulting damage, they could be overlayed for the exploration of possible relationships.

A data warehouse is based on a multidimensional data model which views data in the form of a data cube [1]. A data cube allows data to be modelled and viewed in multiple dimensions and is typically implemented by adopting a star schema model, where the data warehouse contains a large central table called *fact table* which is related with a set of smaller tables called *dimensional tables*, e.g. *quake(q_id, type)*, *time(hour, day, month, year, decade)*. Fact table contains measures (such as *number of earthquakes*) and keys to each of the related dimension tables.

Further to the operations of roll-up and drill-down described above, typical data cube operations are also the operations of *slice* and *dice* for selecting parts of a data cube by imposing conditions on one or more cube dimensions, respectively, and *pivot* which provides alternative presentations of the data to the user.

2.3 Data mining

The integration of data analysis and mining techniques into the SDMMS ultimately aims to the discovery of interesting, implicit and previously unknown knowledge. We study the integration of three basic techniques for this purpose: methods for finding association rules, clustering algorithms and classification methods. Recently, there have been proposals that expand the application of knowledge discovery methods on multi-dimensional data [9, 10].

Association rules have are implications of the form $A \Rightarrow B$, $A \subset J$, $B \subset J$ where A, B and J are sets of items and are characterized by two numbers, s, which is called support of the rule and expresses the probability that a transaction in a database contains both A and B, and c, which is called confidence of the rule and expresses the conditional probability that a transaction containing A also contains B. For example, an association rule could be that earthquakes of $M > 5$ occurred in certain regions in Greece at a specific year with confidence 60% and support 30%.

Clustering algorithms [8, 5] group sets of objects into classes of similar objects. Possible applications on earthquake data could be for the purpose of finding densely populated regions according to the Euclidean distance between the epicenters, and, hence, locating regions of high seismicity.

Classification of data is a two-step process [5]. In the first step a classification model is built using a *training data set* consisting of database tuples that it is known to belong in a certain class (or, in many cases, an attribute of the tuples denotes the corresponding class) and a proper supervised learning method, e.g. decision trees [5]. In the second step, this model is used for the classification of tuples not included in the training set. For example, we could classify earthquake data according to magnitude, location of epicenter or their combination.

Visualization techniques can be used either for the purpose of presenting query results or for assisting the user in the formulation of queries and allowing visual feedback to the Knowledge Discovery in Databases (KDD) process. For example, spatial regions can be selected graphically and queries concerning these regions could be subsequently

formulated or the selection of variables on which classification will be performed could be decided after the examination of a parallel coordinate plot [7]. Furthermore, visual interaction tools allow quick and easy inspection of spatiotemporal relationships as well as evaluation and interpretation of data mining results.

Widely used visualization techniques include the usage of geographic maps for the visualization of the spatial and temporal distribution of data attributes, clusters and classes and tree views, scatter plots and various types of charts for the visualization of mining results. Examples of mining results visualization are the use of tree-views for the results of the application of a decision tree learning algorithm and the use of scatter plots for the visualization of relationships between variables of association rules.

3 A SDMMS prototype

Further to the hints presented in previous sections, if we also consider the integration of heterogeneous data sources over the WWW and the batch nature of the database update process, we propose the architecture presented in figure 1. A number of filters perform the operations of cleaning, transforming and homogenizing data from external data sources (e.g. web sites). A filter has to be implemented and added to the SDMMS every time we would like to import into the database a data set from a new, different source. The data load manager performs the operation of loading the filtered data into the database.

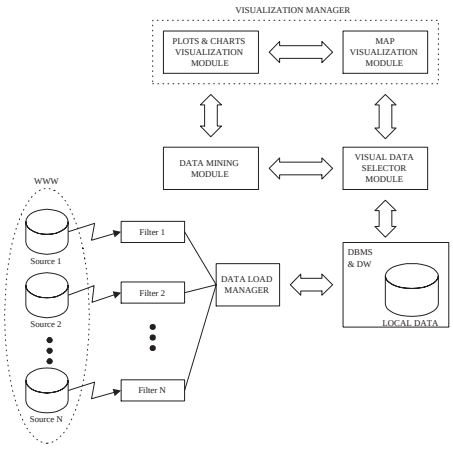

Fig. 1. The proposed architecture of the SDMMS.

Apart from the use of the DW technology, essential parts of the SDMMS architecture are also Data Mining and Visualization modules. The visual data selector module allows the selection of subsets of data, e.g. by visually zooming into an area of interest or by setting constraints on the values of desirable attributes, and the selection of the

abstraction level at which the data are studied, e.g. by performing DW operations like drill-down, roll-up etc. After the user has selected the data of interest, a data visualization technique can be applied by the visualization manager either directly either after performing a mining task on the selected data. The visualization manager is composed of two modules, one performing the generation of plots and charts and one performing visualization of data over maps (figure 2).

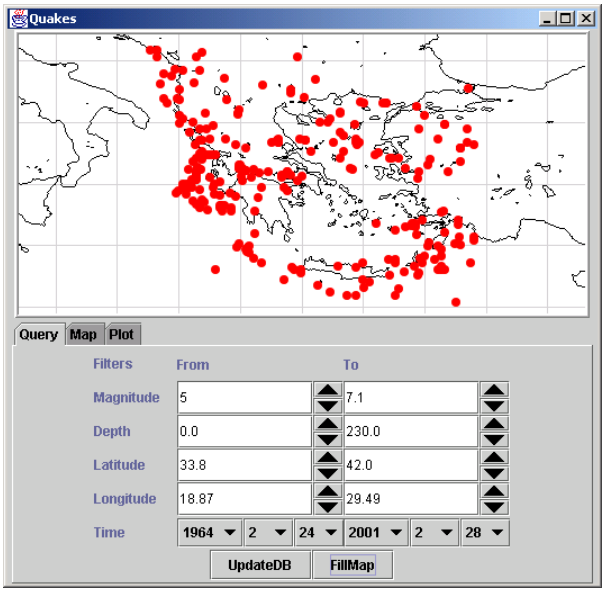

Fig. 2. Screen-shot of our prototype illustrating the spatial distribution of earthquake epicenters for the whole region of Greece for the time period 1964/2/24 to 2001/2/28 that correspond to $M_L \geq 5$.

In an attempt to outline the functionality of the SDMMS prototype, we could divide the operations supported in three main categories:

- *Spatiotemporal Queries.* The SDMMS provides the capability of performing queries concerning the spatiotemporal as well as traditional one-dimensional properties of seismic data and their relationships. Thus, the user can study seismicity and by isolating regions and time periods of interest, execute nearest neighbor queries (e.g. *"find all epicenters of earthquakes in distance no more than 50Km from Athens"*), find earthquakes occurred in inhabited areas (range queries), calculate distances between epicenters and other geographical entities like schools and heavy industries, frequency of disastrous earthquakes occurrence and other statistics. Input of queries is performed with the assistance of visual controls and the results can be presented in form of tables or charts and over maps.
- *Data Cube Operations.* SDMMS provides ways for dealing with the very large size of the observational data sets by supporting summarized views of data in different

levels of abstraction of the spatial (e.g. province, country, continent), temporal (e.g. month, year, ten year period) or any other dimension characterizing the data.
- *Remote data sources management functionality.* The proposed SDMMS architecture provides a flexible implementation framework. For example, as an option, only summaries of seismic data could be stored locally and in case the user requests a detailed data view which is not available, additional data can be loaded, from the remote (web) source, on demand.
- *Simple Mining Operations.* By means of clustering, classification and association rules, SDMMS provides capabilities of constructing seismic profiles of certain areas and time periods, of discovering regions of similar seismic behavior, of detecting time recurrent phenomena and of relating seismic parameters between themselves and to information like damages, population of areas, proximity of epicenter to cities etc.
- *Phenomena extraction.* The data mining module is also used for the automatic extraction of semantics from stored data, such as the characterization of the mainshock and possible intensive aftershocks in shock sequences (see figure 3, for example, where the local magnitudes[1] of a shock sequence are depicted). Moreover, one could search for similar sequences of earthquake magnitudes in time or in space, i.e. time-patterns that occurred in several places or space-patterns that occurred in several time periods.

Furthermore, by providing links between them, the two visualization modules can be used in parallel. For example, rules could be presented in the form of a chart in one window, while, in another window, a map could be showing the spatial distribution of the data for which these rules hold true.

Currently, the prototype is at an early stage of development. Until this moment only one data set, concerning earthquakes of Greece during the time period 1964-2000[2], has been integrated into the database. The user can automatically check if new data have been posted on the web site and update the database. The underlying database technology is currently an RDBMS (Oracle 8i). Queries are formulated using visual controls for constraining the allowed values of both the alphanumeric and spatiotemporal attributes of data. Basic map functionality (zoom in/out and shift operations) is also provided, while pairs of attributes of the earthquakes whose epicenters are shown on the map can be further explored by means of 2d plots.

Next stages of development mainly include the implementation of more sophisticated visualization techniques, of the mining module, of the DW and of more filters for the integration of other web data sources. We have selected to implement the prototype within a modern commercial DBMS like Oracle, where special data types and predicates for spatiotemporal data can be defined and implemented, allowing the full integration of the spatiotemporal models described in section 2 into the database system. Furthermore, we plan to migrate to version 9i as soon as becomes available, which

[1] An earthquake magnitude scale based on measurements of the amplitude of earthquake waves recorded on a standard Wood-Anderson type seismograph at a distance of less than 600 kilometers.

[2] It is publicly available through the web site of the Institute of Geodynamics, National Observatory of Athens, Greece [http://www.gein.noa.gr].

provides an API for extensible indexing, meaning that we can construct spatiotemporal indexes of our choice, for any new data types we define, and achieve, consequently, superior performance when querying spatiotemporal attributes. The data warehouse will also be implemented by taking advantage of the ready to use functionality provided within the Oracle DBMS including support for constructing data cubes, for performing aggregation queries etc. Finally, as data analysis tools, we plan to use combination of the DBMS's of-the-self tools and of our own implementations of mining algorithms of our choice.

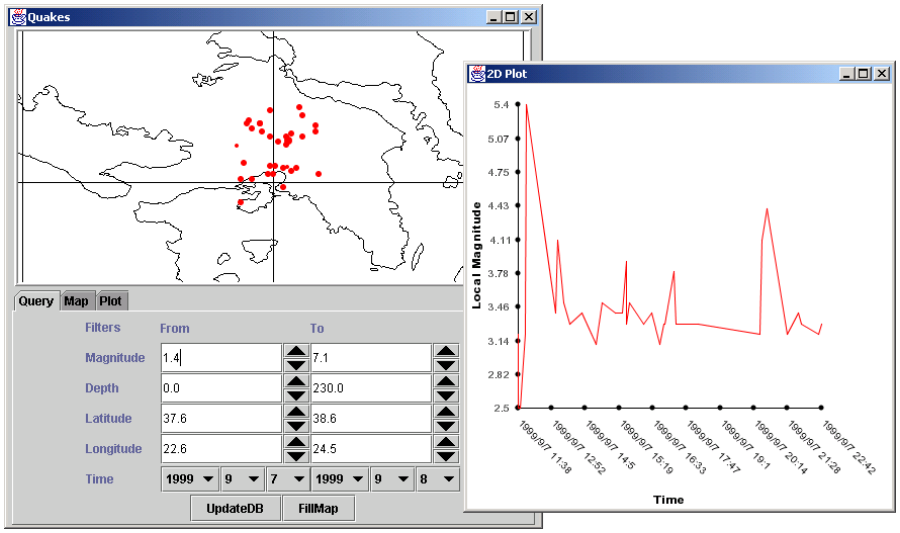

Fig. 3. A time sequence of 33 shocks occurred on the 7th of September 1999 in Athens, Greece. The greatest peak refers to the mainshock of local magnitude $M_L = 5.4$, while it is clear that intensive aftershocks followed.

4 Related work

Taking into consideration the availability of massive spatiotemporal data sets, the need for high level management, analysis and mining has already been recognized by several researchers as shown by the development of several research prototypes. In the sequel, we briefly describe the purpose and functionality of the most important, to our knowledge so far, that have appeared in the literature. We distinguish two categories of interest. The first includes prototypes that perform one or more of the operations of managing, mining and visualizing either spatial or spatiotemporal data, while the second includes analogous prototypes - applications for scientific and geophysical data.

Geo-miner [6] is a data mining system for spatial data that includes modules for mining (rules, classifications, clusters), for spatial data cube construction and for spa-

tial OLAP. These modules are accompanied with a geo-mining query language and visualization tools.

In [2], the building of an integrated KDD environment for spatial data is proposed, based on the prototypes Kepler, for data mining, and Descartes, for interactive visual analysis. This work is primarily focused on visualization techniques that increase the effectiveness of the KDD process by means of sophisticated user interaction.

In [3], five case studies are presented in order to bring up critical issues and to show the potential contribution of applying data mining techniques for scientific data while application prototypes for several scientific fields (SKICAT for determining whether observed sky objects are stars or galaxies, JarTool for searching for features of interest on the surface of planets, a system for mining bio-sequence databases, Quakefinder for detecting tectonic activity from satellite data and CONQUEST for the analysis of atmospheric data) are presented. This study is mainly concerned with the issues of dealing with the very large size of observational data sets and of how KDD techniques can be used for extracting phenomena on high conceptual level from the low-level data.

Finally, commonGIS [11] is a web-based system for the visualization and analysis of spatially-related statistical data based on the previously mentioned Descartes prototype and on a geo-spatial database. CommonGIS has been used for the construction of a database [3] of earthquake events registered within the European Countries between 500 AC and 1984 DC.

We intend to combine some interesting functionalities of the above systems and integrate them into the SDMMS prototype. What most distinguishes our SDMMS prototype is that it proposes a integrated environment for managing and mining spatial data and for the exploitation of heterogeneous data sources.

5 Conclusion

In this paper, we have proposed a novel SDMMS architecture, described its functionality and outlined the potential benefits, by providing extended examples, for potential users: researchers of geophysical sciences, key personnel in public administration as well as people who are just interested on querying or viewing seismic data. Main issues discussed include spatiotemporal concepts necessary for the modelling of seismic activity and for efficient storage and retrieval of seismic data, KDD and DW technologies for dealing with the very large amount of available seismic parameters observations and for their effective analysis, and visualization techniques that empower the user to fully exploit the capabilities of the SDMMS by linking it with KDD operations.

Additionally, we have taken into consideration the reported experiences from the development of several prototypes for mining, managing and visualizing spatial, spatiotemporal and geophysical data and we intend to make use of them in a systematic way in the next stages of development of the SDMMS prototype. As a further step, we propose a framework based on DW technology for the integration of publicly available seismic data sources into the SDMMS prototype.

A web interface of the SDMMS prototype, a generator of tectonic activity scenarios and simulator in the line of the GSTD generator for spatiotemporal data [15] and a

[3] It is publicly accessible at [http://commongis.jrc.it/sw/commongis_first/equakes1.html]

study of how other KDD techniques can be applied to seismic data are directions that could be followed for the further expanding of SDMMS prototype capabilities.

References

1. S. Agarwal, R. Agrawal, P. Deshpande, A. Gupta, J. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. In *Proceedings of the 22nd International Conference on Very Large Databases, VLDB*, pages 506–521, Bombay, India, September 1996.
2. G. Andrienko and N. Andrienko. Knowledge-based visualization to support spatial data mining. In *Proceedings of the 3rd Symposium on Intelligent Data Analysis*, pages 149–160, Amsterdam, The Netherlands, August 1999.
3. U. Fayyad, D. Haussler, and P. Stolorz. KDD for science data analysis: Issues and examples. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 50–56, Portland, Oregon, USA, 1996.
4. R. Guting, M. Bohlen, M. Erwig, C. Jensen, N. Lorentzos, M. Schneiderand, and M. Vazirgiannis. A foundation for representing and quering moving objects. *ACM Transactions on Database Systems (TODS)*, 25(1):1–42, March 2000.
5. J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.
6. J. Han, K. Koperski, and N. Stefanovic. GeoMiner: a system prototype for spatial data mining. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 553–556, Tucson, Arizona, USA, May 1997.
7. A. Inselberg. Multidimensional detective. In *Proceedings of IEEE Visualization Conference*, pages 100–107, Phoenix, AZ, October 1997.
8. A. Jain, M. Murty, and P. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, Volume 31(3):264–323, September 1999.
9. K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Proceedings of the 4th International Symposium on Large in Spatial Databases, SSD*, pages 47–66, Portland, Maine, August 1995.
10. K. Koperski, J. Han, and J. Adhikary. Mining knowledge in geographical data. *Communications of the ACM*, 26(1):65–74, March 1998.
11. U. Kretschmer and E. Roccatagliata. CommonGIS: a European project for an easy access to geo-data. In *Proceedings of the 2nd European GIS Education Seminar, EUGISES*, Budapest, Hungary, September 2000.
12. M. Nascimento and J. Silva. Towards Historical R-Trees. In *Proceedings of ACM symposium on Applied Computing*, pages 235–240, Atlanta, GA, USA, February/March 1998.
13. D. Pfoser, C. Jensen, and Y. Theodoridis. Novel approaches in query processing for moving object trajectories. In *Proceedings of the 26th International Conference on Very Large Data Bases, VLDB 2000*, pages 395–406, Cairo, Egypt, September 2000.
14. D. Pfoser and N. Tryfona. Requirements, definitions, and notations for spatiotemporal application environments. In *Proceedings of the 6th international symposium on Advances in Geographic Information Systems, ACM-GIS '98*, pages 124–130, Washington, DC, USA, November 1998.
15. Y. Theodoridis and M. Nascimento. Spatiotemporal datasets on the WWW. *SIGMOD Record*, 29(3):39–43, September 2000.
16. Y. Theodoridis, M. Vazirgiannis, and T. Sellis. Spatio-temporal indexing for large multimedia applications. In *Proceeding of the 3rd IEEE Conference on Multimedia Computing and Systems*, pages 441–448, Hiroshima, Japan, 1996.