Seismic data management and mining systems – An Overview

Gerasimos Marketos and Yannis Theodoridis

Department of Informatics, University of Piraeus

Abstract

We present the architecture of a so-called Seismic Data Management and Mining System (SDMMS) for quick and easy data collection, processing, and visualization. The proposed SDMMS architecture includes, among others, a seismological database for efficient and effective querying and a seismological data warehouse for OLAP analysis and data mining. We provide template schemes for these two components as well as examples of their functionality. We also provide a survey of existing operational or prototype SDMMS.
Keywords: seismological databases, data warehousing, data mining

1. Introduction

For centuries, humans have been feeling, recording and studying earthquake phenomena. Taking into account that at least one earthquake of magnitude M<3 (M>3) occurs every one second (every ten minutes, respectively) worldwide, the seismic data collection is huge and, unfortunately, rapidly increasing. Scientists record this information in order to describe and study tectonic activity. Tectonic activity can be described by recording attributes,

Πανεπιστήμιο Πειραιώς – University of Piraeus
Τιμητικός τόμος Ομότιμου Καθηγητού Αντωνίου Χ. Παναγιωτόπουλου (2006) 629–646
Volume of essays in honour of Professor Antonios C. Panayotopoulos (2006) 629–646

such as geographic information (epicenter location and disaster areas), magnitude, depth, etc.

On the other hand, computer engineers specialized in the area of Information & Knowledge Management find an invaluable «data treasure», which they can process and analyze helping in the discovery of knowledge from this data. Recently, many applications that can manage and analyze seismological or, in general, geophysical data have been proposed in the literature (Han et al., 1997; Andrienko and Andrienko, 1999; Theodoridis, 2003). In general, the collaboration between the data mining community and physical scientists was recently initiated (Behnke and Dobinson, 2000).

Desirable components of a so-called Seismic Data Management and Mining System (SDMMS) include tools for quick and easy data exploration and inspection, algorithms for generating historic profiles of 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 (e.g. topological and climatological) maps for the presentation of data to the user and, topline, visualization components supporting sophisticated user interaction.

The rest of the paper is organized as follows. In the Section II we sketch a desired SDMMS architecture, including its database and data warehouse design. Section III presents online analytical processing (OLAP) and data mining functionality a SDMMS could offer. In Section IV, we survey and compare proposed systems and tools found in the literature for the management of seismological or, in general, geophysical data. Finally, Section V concludes the paper.

2. The Architecture of a Seismic Data Management and Mining System

Earthquake phenomena are instantly recorded by a number of organizations (e.g. Institutes of Geodynamics and Schools of Physics) worldwide. The architecture of a SDMMS might allow for the integration of several remote sources. The aim is to

collect and analyse the most accurate seismic data among different sources. Obviously, some sources provide data about the same earthquakes though with slight differences in their details (the magnitude or the exact time of the recorded earthquake). SDMMS should be able to integrate the remote sources in a proper way by refining and homogenizing raw data.

Collected data can be stored in a local database and/or a data warehouse (for simple querying and analysis for decision making, respectively). Data within the database is dynamic and detailed, while that within the data warehouse is static and summarized. The modifications of the former are continuous, while the latter are subjected to periodical updates.

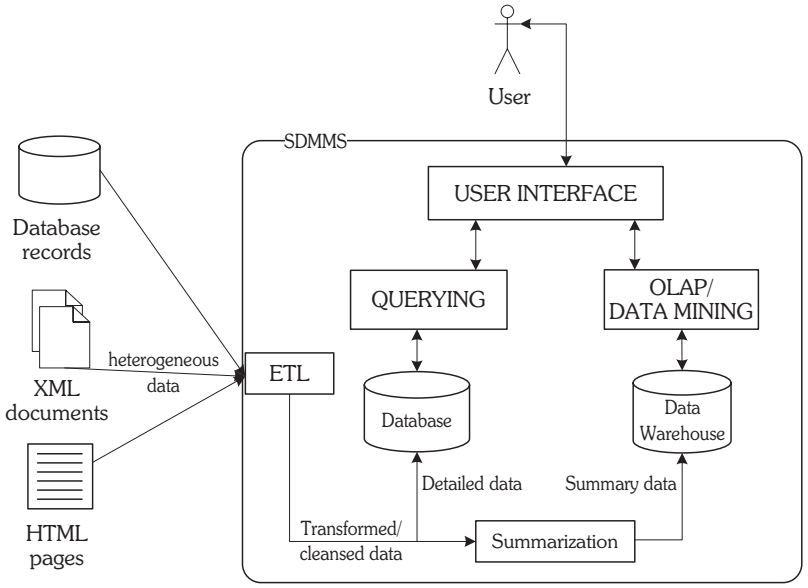

Fig. 1. The general SDMMS architecture

Figure 1 presents an abstract architecture that serves the task of collecting data from several sources around the world and storing them in a local repository (database and/or data warehouse). A mediator is responsible for the management of the process from the extraction of data from their sources until their

load into the local repository, the so-called Extract-Transform-Load (ETL) approach.

In the following paragraphs, we present efficient design proposals for the core components of SDMMS architecture, namely the database and the data warehouse.

2.1 SDMMS database

Remote sources provide SDMMS with a variety of seismological information to be stored in the local database. Figure 2 illustrates the relational design of a local database for SDMMS purposes.

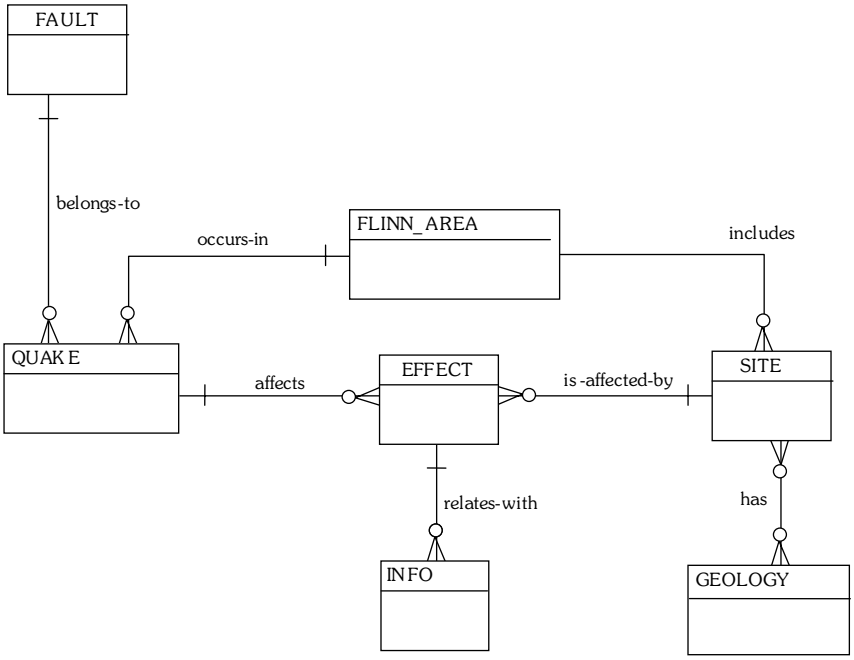

Fig. 2. Relational design of a seismological database for SDMMS purposes

The minimum information required to describe an earthquake event includes timestamp of its appearance, location (in latitude / longitude coordinates) and depth (QUAKE). Just this information is not adequate for user-friendly querying and further data analysis

as one wishes to know more about the geographical areas where an earthquake occurred. For this purpose, the addition of FLINN_AREA assists on the geographical positioning of both the earthquake epicenter and the affected sites using the Flinn & Engdahl geographical terminology (Young et al., 1996) that partitions world in disjoint polygons. Moreover, FAULTS includes details about the seismogenic fault related with an earthquake (name of the fault, its characterization, strike, slip and rake of planes, etc.), extracted from bibliography, e.g. (Kiratzi and Louvari, 2003).

SITE stores demographical and other information about the primitive administrative partitions of a country (e.g. counties or municipalities). GEOLOGY describes the geological morphology of a site so that we can discover how the different morphological classes are affected by earthquakes. EFFECT records macroseismic intensity observed at a site as a result of an earthquake. Other attributes of this relation might include the epicentral and hypocentral distance and the azimuth (the angle between the site-epicenter line and the line of North). Finally, INFO includes complementary multimedia material, such as pictures, audio/video descriptions, references etc. about the earthquake effects.

2.2 SDMMS data warehouse

A data warehouse (DW) is defined as a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management decision making process (Inmon, 1996). DW is usually based on a multi-dimensional data model, which views data in the form of a data cube (Agarwal et al, 1996). A data cube allows data to be modeled and viewed in multiple dimensions and is typically implemented by adopting a star (or snowflake) schema model, where the DW consists of a fact table (schematically, at the center of the star) surrounded by a set of dimensional tables related with the fact table. For SDMMS purposes, dimensional tables maintain information (e.g. hierarchies) about dimensions, including magnitude, intensity, geography, time dimension, etc. while the fact table contains measures on seismological data, such as the number of earthquakes,

minimum/maximum depth, etc. as well as keys to related dimensional tables (Figure 3). Especially for SDMMS, where geography is a key issue and is involved in dimensions and/or measures, spatial data warehouses are of great interest (Stefanovic et al, 2000).

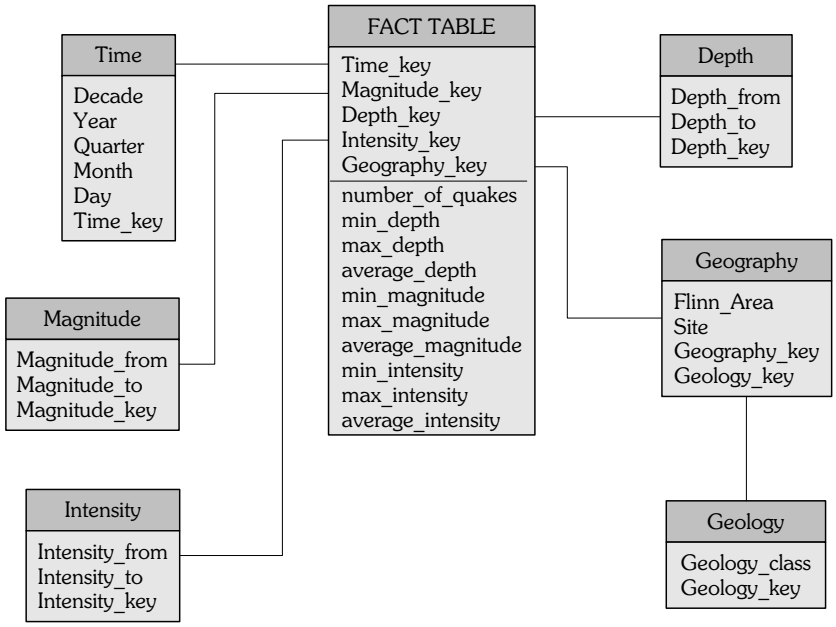

Fig. 3. A spatial data warehouse for SDMMS purposes

Dimension time consists of a hierarchy that represents time periods in which an earthquake happened. Dimensions magnitude, intensity and depth do not consist of hierarchies but of intervals. They represent classes of magnitude, intensity and depth so that we can categorize the earthquake phenomena. Dimensions geography and geology represent the geographical area in which an earthquake happened and the geological morphology of this area.

In the following Section we present examples of operations that illustrate the usefulness of a data warehouse following the scheme of Figure 3.

3. OLAP analysis and data mining

Traditional Database Management Systems (DBMS) are known as operational databases or OLTP (on-line transaction processing) systems as they support the daily needs of Information Systems for storage and retrieval. They support three main operations (insertions, updates and deletions) that can be formalized and executed against a DBMS using a Structured Query Language (SQL).

However, Information Systems are not responsible only for data storage and retrieval but also for supporting decision making. As already mentioned, maintaining summary data in a local data warehouse can be used for data analysis purposes. Two popular techniques for analyzing data and interpreting their meaning are OLAP analysis and data mining.

An important aspect in decision making is the level of details that the decision maker needs. Middle and upper management make complex and important decisions and therefore detailed data can not satisfy these requirements. Summarized data and hidden knowledge acquiring from the stored data, can lead to better decisions. Similarly, summarized seismological data are more useful to scientists because they can study them from a higher level and search them for hidden, previously unknown knowledge.

In the following paragraphs, we present OLAP analysis and data mining techniques for extracting useful conclusions about seismological data stored in a SDMMS.

3.1 OLAP analysis

Additional to (naive or advanced) database queries on detailed seismological data, a data warehouse approach utilizes on-line analytical processing (OLAP). 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 historical seismic profile, i.e. the ten most destructive quakes in the past twenty years, over Europe, and, moreover, he/she can easily view the same

information over Greece (more detailed view, formally a drill-down operation) or worldwide (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 overlaid for the exploration of possible relationships, such as finding regions of high, though non-destructive, seismicity and vice versa.

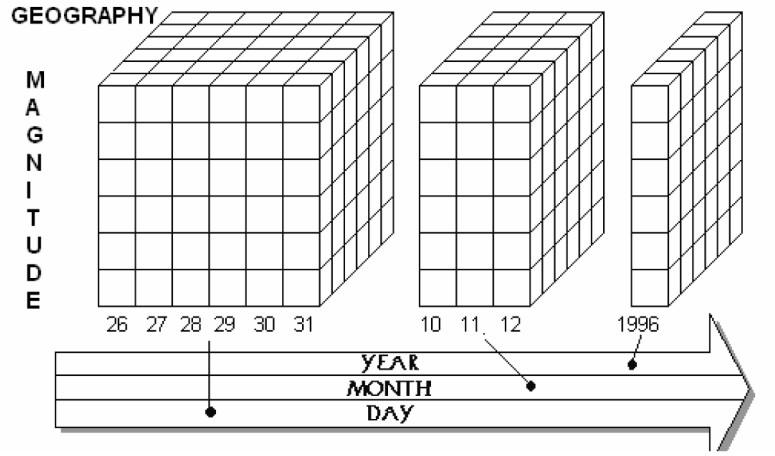

Fig. 4. Moving from a detailed to a summarized view (roll-up) and vice-versa (drill-down)

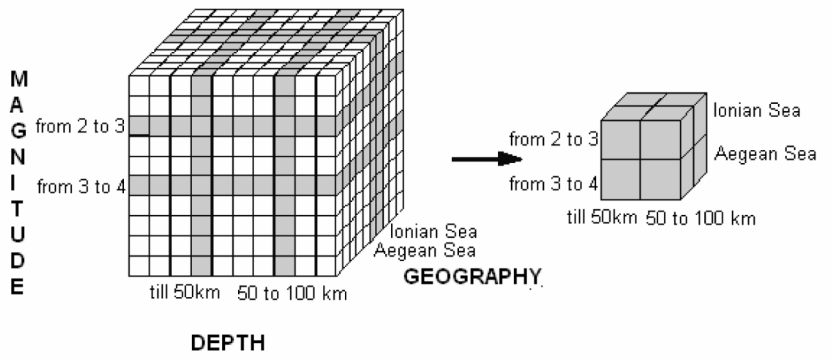

Fig. 5. Selecting parts of a cube by filtering a single (slice) or multiple dimensions (dice).

Further to roll-up and drill-down operations described above, typical data cube operations include slice and dice, for selecting parts of a data cube by imposing conditions on a single or multiple cube dimensions, respectively.

Another important issue in data warehousing is the physical representation of a data warehouse (cube). ROLAP (Relational OLAP) and MOLAP (Multidimensional OLAP) are the two proposed principal models. The advantage of ROLAP is that it can handle large volumes of data (as relational databases can do) as these are stored in relational tables. On the other hand, MOLAP is much faster as it uses specialized data structures instead of relational tables. As a result, in MOLAP model, main memory is extensively used for the various operations.

3.2 Data mining

Integrating data analysis and mining techniques into an SDMMS ultimately aims to the discovery of interesting, implicit and previously unknown knowledge. Figure 6 presents the Knowledge Discovery in Databases (KDD) process with the necessary 'filters' that information stored in a data warehouse passes until useful, possibly hidden knowledge is extracted.

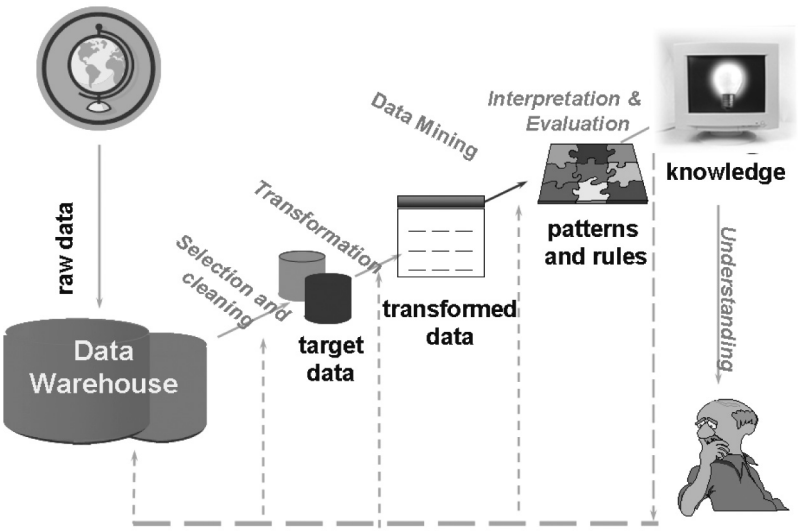

Fig. 6. The KDD process

Examples of useful patterns found through KDD process include clustering of information (e.g. shocks occurred closely in space and/or time), classification of phenomena with respect to area and epicenter, detecting phenomena semantics by using pattern finding techniques (e.g. characterizing the main shock and possible intensive aftershocks in shock sequences, measuring the similarity of shock sequences, according to a similarity measure specified by the domain expert, etc.).

In the following, we study the integration of three basic techniques for this purpose: methods for finding association rules, algorithms for data clustering and classification techniques. Recently, there have been proposals that expand the application of knowledge discovery methods on multi-dimensional data (Koperski and Han, 1995; Koperski et al, 1998).

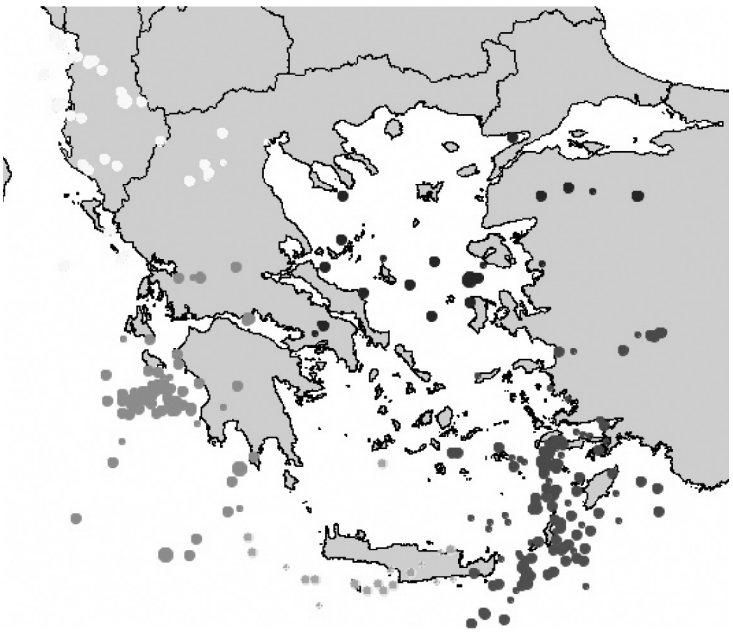

Fig. 7. Discovering clusters of earthquake epicenters (extracted from (Theodoridis, 2003))

Clustering

Data clustering algorithms (Jain and Murty, 1999) group sets of objects into classes of similar objects. Thus, the behavior of groups can be studied instead of that of individuals. Applications on seismic 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 seismic frequency or dividing the area of a country into a set of seismicity zones (e.g. low / medium / high seismic load).

Data classification

Data classification is a two-step process (Han and Kamber, 2000). In the first step a classification model is built using a training data set consisting of database records that are known to belong in a certain class and a proper supervised learning method, e.g. decision trees or neural networks. In case of decision trees, for example, the model consists of a tree of «if» statements leading to a label denoting the class the record belongs in.

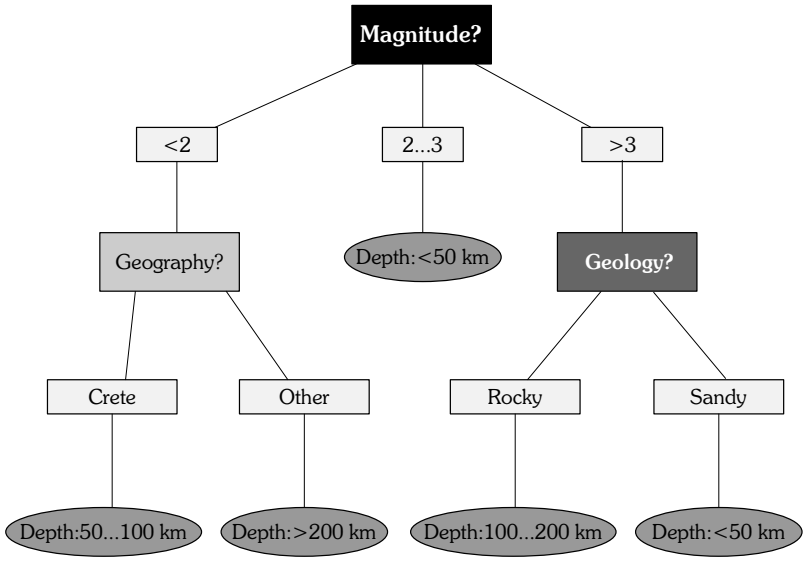

Fig. 8. A decision tree for seismological data

The decision tree illustrated in Figure 8 «predicts» the depth of a future earthquake given the magnitude and details about the area in which it will happen. Obviously, it is not a kind of earthquake prediction since it assumes that we know some details of the earthquake before it happens. However, the above tree could uncover some hidden relationships among the characteristics of earthquakes.

4. Prototype systems and tools: a survey

4.1 Geo-Miner

Han et al. (1997) have developed Geo-Miner, a system that supports knowledge discovery from spatial data. Geo-Miner consists of a number of modules including a spatial data cube construction module, spatial OLAP module, and spatial data mining modules. Extending Spatial SQL, GMQL (Geo-Mining Query Language) has been designed and implemented for spatial data mining. Figure 9 illustrates the functionality of the system through a representative screenshot.

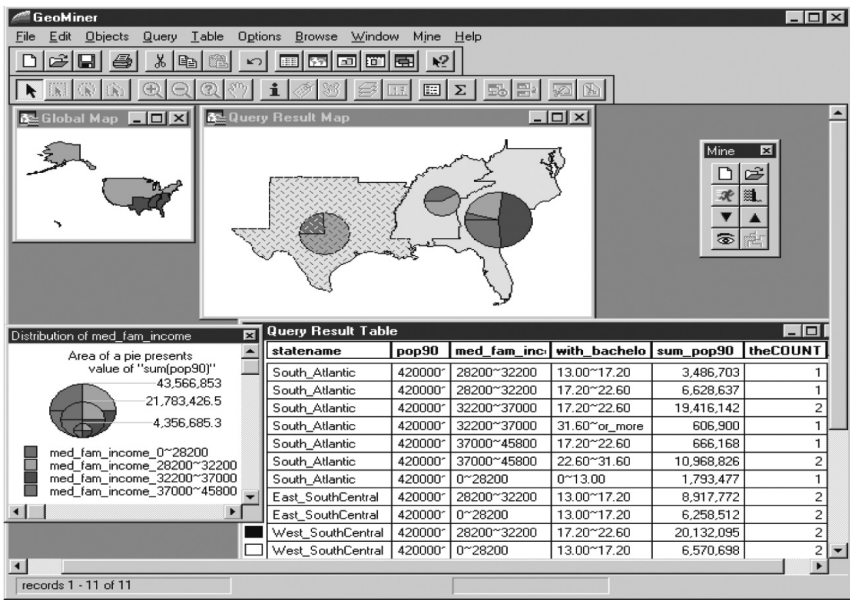

Fig. 9. Geo-Miner

4.2 Descartes / Kepler

Andrienko and Andrienko (1999) have proposed an integrated environment (Descartes/Kepler) where data mining and visualization techniques are used to analyze spatial data. Their aim is to integrate traditional data mining tools with cartographic visualization tools so that the users can view both source data and results produced by the data mining process. Figure 10 illustrates a composite screenshot of the tool with maps and charts visualization.

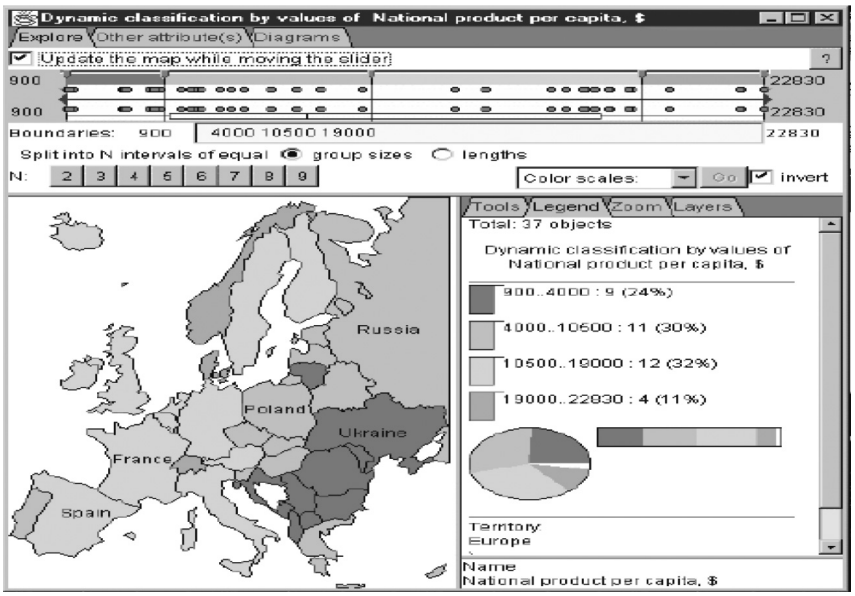

Fig. 10. Descartes / Kepler

4.3 GEODE

Geo-Data Explorer (GEODE) is an ambitious and high promising application developed by the USGS for providing users with geographically referenced data. The project aims in developing a portal which will provide real-time data and will support data analysis independently from special hardware, software and training (Levine and Schultz, 2002). Figure 11 illustrates the functionality of the system through a representative screenshot.

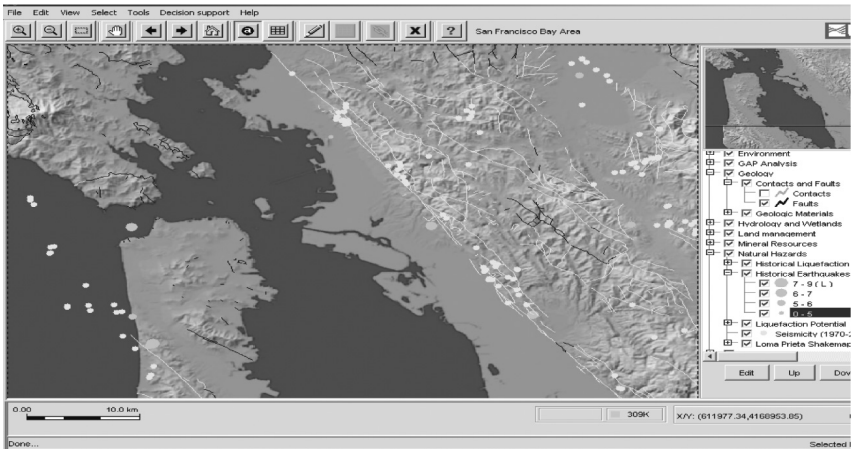

Fig. 11. GEODE functionality

4.4 Seismo-Surfer

Last but not least, Seismo-Surfer is a tool for collecting, querying, and mining seismological data following the SDMMS concept. Its database is automatically updated from remote sources, querying on different earthquake parameters is allowed, while data analysis for extracting useful information is limited to a data clustering algorithm. Querying and mining results are graphically presented via maps and charts.

Seismo-Surfer architecture in general follows the SDMMS architecture illustrated in Figure 1. A number of filters «clean» and homogenize the data (mainly concerning duplicate entries), which are available from remote sources and cleansed data are stored in the local database. Users interact with the database via a graphical user interface (called, Query Manager). KDD techniques apply data mining on stored data. Querying and data mining results are presented in graphical mode (maps, charts, etc.).

In its current version, Seismo-Surfer supports links with two remote sources: one at a national level for Greece (GI-NOA) and one worldwide (NEIC-USGS). Querying on earthquake parameters includes variations of spatial queries, such as range, distance, nearest-neighbor and closest-pair queries (Figure 12).

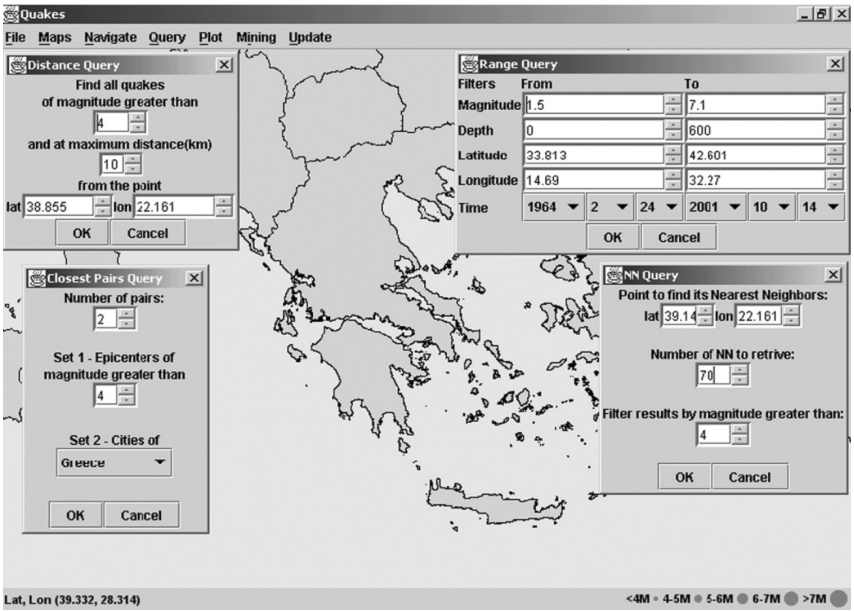

Fig. 12. Querying facilities using Seismo-Surfer

5. Conclusions

In this paper, we outlined the architecture of a so-called Seismic Data Management and Mining System (SDMMS) for quick and easy data collection, processing (generating historic profiles of specific geographic areas and time periods, providing the association of seismic data with other geophysical parameters of interest, etc.), and visualization supporting sophisticated user interaction.

The core components of such a SDMMS architecture include a seismological database (for querying) and a seismological data warehouse (for OLAP analysis and data mining). We provided template schemes for both components as well as examples of their functionality.

Finally, we provided a survey of existing operational or prototype systems following (at a low or high percentage) the proposed SDMMS functionality.

References

Agarwal, S., Agrawal, R., Deshpande, P., Gupta, A., Naughton, J., Ramakrishnan, R., Sarawagi, S. (1996): On the computation of multidimensional aggregates. *In Proceedings of the 22nd International Conference on Very Large Databases*, VLDB'96, pp. 506-521, Bombay, India.

Andrienko, G., Andrienko N. (1999): Knowledge-based visualization to support spatial data mining. *In Proceedings of the 3rd Symposium on Intelligent Data Analysis*, IDA'99, pp. 149-160, Amsterdam, the Netherlands.

Behnke, J., Dobinson, E. (2000): NASA Workshop on Issues in the Application of Data Mining to Scientific Data, *ACM SIGKDD Explorations Newsletter*, 2(1), pp. 70-79.

GI-NOA: Earthquake Catalog. Available at http://www.gein.noa.gr/services/cat.html (accessed 26 January 2005).

Han, J., Kamber, M. (2000): *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Han, J., Koperski K., Stefanovic N. (1997): GeoMiner: A System Prototype for Spatial Data Mining. *In Proceedings of ACM SIGMOD International Conference on Management of Data*, SIGMOD'97, pp. 553-556, Tucson, AZ, USA.

Inmon, W. (1996): *Building the Data Warehouse*, 2nd ed., John Wiley.

Jain, A., Murty, M., Flynn, P. (1999): Data Clustering: A Review. *ACM Computing Surveys*, 31(3), pp. 264-323.

Kiratzi, A., Louvari, E. (2003): Focal Mechanisms of Shallow Earthquakes in the Aegean Sea and the Surrounding Lands Determined by Waveform Modeling: A New Database. *Journal of Geodynamics*, 36, pp. 251-274.

Koperski K., J. Han J. (1995): Discovery of Spatial Association Rules in Geographic Information Databases. *In Proceedings of the 4th International Symposium on Large in Spatial Databases*, SSD'95, pp. 47-66, Portland, Maine, USA.

Koperski, K., Han, J., Adhikary, J. (1998): Mining Knowledge in Geographical Data. *Communications of the ACM*, 26(1), pp. 65-74.

Levine, M., Schultz, A. (2002): GEODE (Geo-Data Explorer) - A U.S. Geological Survey Application for Data Retrieval, Display, and Analysis through the Internet , U.S. Geological Survey, Fact Sheet 132-01, Online Version 1.0. Available at http://pubs.usgs.gov/fs/fs132-01/ (accessed 26 January 2005).

NEIC-USGS: Earthquake Search. http://neic.usgs.gov/neis/epic/epic_global.html (accessed 26 January 2005).

Stefanovic, N., Han, J., Koperski, K. (2000): Object-Based Selective Materialization for Efficient Implementation of Spatial Data Cubes. IEEE *Transactions on Knowledge and Data Engineering*, 12(6), pp. 938-958.

Theodoridis, Y. (2003): Seismo-Surfer: A prototype for collecting, querying and mining seismic data. *In Advances in Informatics - Post Proceedings of the 8th Panellenic Conference in Informatics*, pp. 159-171, LNCS #2563, Springer - Verlag, Berlin.

Young, J.B., Presgrave, B.W., Aichele, H., Wiens, D.A. Flinn, E.A. (1996): The Flinn-Engdahl Regionaligation Scheme: The 1995 Revision. *Physics of the Earth and Planetary Interiors*, 1996, pp. 223-297.