Δελτίο της Ελληνικής Γεωλογικής Εταιρίας τομ. XXXVI, 2004
Πρακτικά 10ου Διεθνούς Συνεδρίου, Θεσ/νίκη Απρίλιος 2004

Bulletin of the Geological Society of Greece vol. XXXVI, 2004
Proceedings of the 10th International Congress, Thessaloniki, April 2004

A TOOL FOR COLLECTING, QUERYING AND MINING MACROSEISMIC DATA

Kalogeras I. S.[1], Marketos G.[2], and Theodoridis Y.[2]

[1] *Geodynamic Institute, National Observatory of Athens, P.O.Box 20048, 11810 Athens, Greece, i.kalog@gein.noa.gr*

[2] *Department of Informatics, University of Piraeus, 18534 Piraeus, Greece, ytheod@unipi.gr*

ABSTRACT

SEISMO-SURFER is a tool for collecting, querying and mining seismic data being developed in Java programming language using Oracle database system. The objective is to combine recent research trends and results in the fields of spatial and spatio-temporal databases, data warehouses and data mining, as well as well established visualization techniques for geographical information. The database of the tool is automatically updated from remote sources while existing possibilities allow the querying on different earthquakes parameters, the analysis of the data for extraction of useful information and the graphical representation of the results via maps, charts etc.

In the present work, we extend SEISMO-SURFER to include macroseismic data collected by the Geodynamic Institute and filled in a relative database. More specifically, the seismic parameters of the strong earthquakes, stored into the SEISMO-SURFER database, are linked to the macroseismic intensities observed at different sites. Administrative information for each site, local surface geology, tectonic lines, damage photographs and detailed descriptions from newspapers are also included.

University of Piraeus and Geodynamic Institute are working together to continuously update and develop SEISMO-SURFER, concerning the data included, the variety of parameters stored and the mining algorithms supported for exploiting knowledge.

1 INTRODUCTION

Aiming to develop a prototype software in order to combine recent research trends and results within the areas of spatial and spatio-temporal databases, data warehouses and extraction of new knowledge from large databases (data mining), lead to the development of the SEISMO-SURFER tool (Theodoridis 2003). SEISMO-SURFER is an example of cooperation between scientists of Informatics (especially those who involved in the information and knowledge management) and of Geophysics.

As seismological data are multidimensional, they need to be stored and recovered by special techniques, more complex compared to those used for the traditional alphanumerical data. Under this point of view, spatial entities referred to temporal periods or temporal moments referred to layers of geographical information are under investigation within the frame of Database Management Systems. Furthermore data warehouse techniques are used in order to unify different sources of seismological data (available through Internet, for example). A user might ask information about the most destructive earthquakes in Europe during the last 20 years, or to limit his/her question in Greece only (drill-down operation) or to extent it world widely (roll-up operation). On the other hand, different layers of thematic information can be included, like geological maps, tectonic maps, population maps etc, in order for the user to search for possible relations between the grade of damage and the epicentral distance or the distance of the damaged cities from the seismogenic fault, or between the damage and the dominant geology etc.

From the above mentioned simple examples, SEISMO-SURFER can be useful to a wide range of users, including scientists of Informatics and Seismology, government officers, even students and simple citizens. Under this point of view the tool should be continuously updated and improved con-

cerning the records, the parameters included, the data-mining techniques and the visualization techniques supported.

2 ARCHITECTURE AND OPERATION OF THE TOOL

The system architecture is illustrated in Figure 1. A number of filters "clean" and homogenize the data (concerning mainly the double entries), which are available from the remote sources and Data Load Manager loads the so resulted data in the Local Database. Users interact with the database via the Query Manager. The Data Mining Module applies knowledge discovery techniques on stored data. Querying and data mining results are presented in graphical mode (maps, charts, etc.) via the Visualization Manager.

The operation of the SEISMO-SURFER tool is classified as follows:

1. *Remote data sources management*: The local database is automatically updated using up-to-date information found in remote sources. In the current implementation phase, two sources are linked in the system: one for phenomena in Greece (source: Geodynamic Institute) and one world-wide (source: US Geological Survey). As an indication, ~~all~~ earthquakes above 4 Richter occurred in Greece since 1964 have been integrated in the SEISMO-SURFER database by using the web site of the Geodynamic Institute, while the catalogue of Comninakis and Papazachos (1986) is used for the integration of the earthquakes in Greece for the period 1901 - 1963.

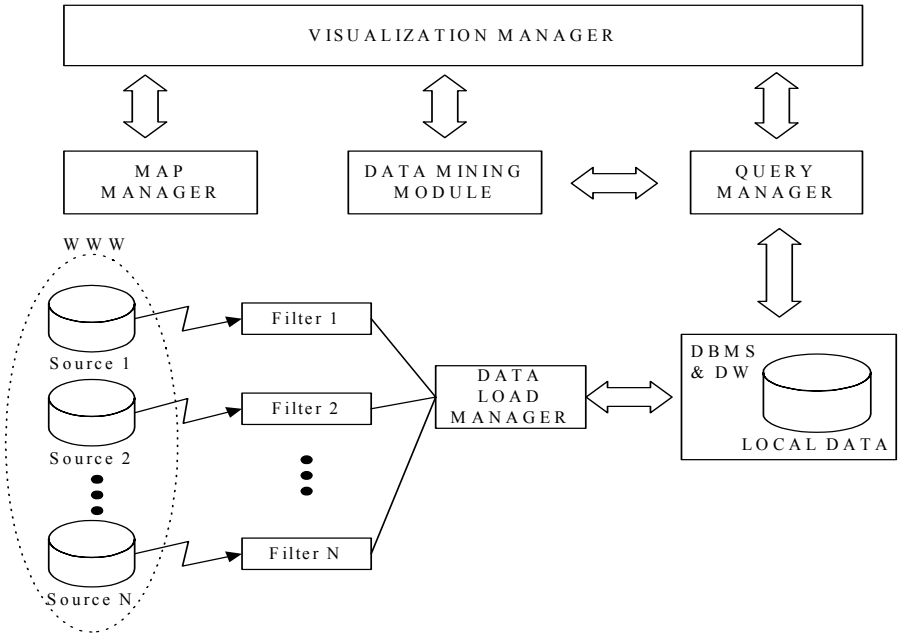

Figure 1. The SEISMO-SURFER architecture.

2. *Querying on seismic parameters of earthquakes:* For example, one could ask about epicenters of earthquakes in distance no more than 50km from Athens, the nearest epicenter with respect to a given point on the map, highly populated cities located close to strong earthquakes etc. (Figure 2).

3. *Data Warehouse operations:* Due to the potentially huge amount of information related to earthquake phenomena, summary data views could equally address user needs. Implementing SEISMO-SURFER as a data warehouse offers the possibility to store locally summary information only (e.g. total number of shocks, average magnitude), at different levels of detail in spatial (country, continent etc.) and/or temporal dimension (month, year, century etc.). Should the details of a single event be asked for retrieval, the remote data sources would be visited.

4. *Data mining operations:* Finding clusters of information (e.g. shocks occurred closely in space and/or time) or classifying phenomena with respect to area and epicenter could be useful tools for earthquake data analysis. Spatial and time-series data mining provide a variety of techniques towards this goal.
5. *Detecting phenomena semantics:* Examples include the characterization of the main shock and possible intensive aftershocks in shock sequences, using pattern finding techniques, the similarity of shock sequences, according to a similarity measure specified by the domain expert, etc.

3 PRESENT SITUATION

SEISMO-SURFER has been developed under Java language with the database running under Oracle environment, aiming to support the possibility for the tool to be web-based, to support different geographical coordination systems, to support spatial querying using specific indexes in order for the searching time to be reduced and to improve data mining– already prepared algorithms can be embedded after small modifications.

Spatial and spatio-temporal queries have been incorporated using the R-tree indexing technique provided by Oracle, including the range query (find the epicenters within an area), the distance query (find the epicenters within distance x from a specific point on the map), the nearest neighbor query (epicenters closest to a point on the map) and the closest pair query (epicenters closest to highly populated cities of Greece). Figure 2. is a mixture of spatio-temporal query screenshots.

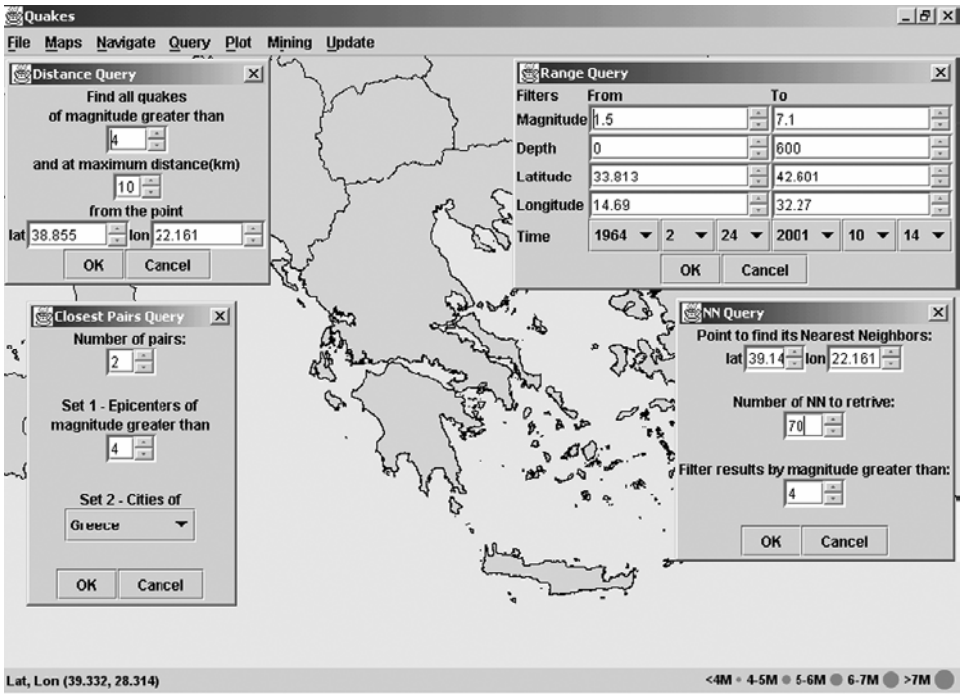

Figure 2. SEISMO-SURFER screenshots for spatio-temporal queries.

For data clustering the popular k-means algorithm has been incorporated using as parameter the number of clusters to be produced. Currently, clustering is based on spatial information only (epicenters).

The tool also includes visualization features like maps, plots and GUI tools that assist the user to the query formulation process and allow viewing the selected or analyzed data in a number of different ways.

4 MACROSEISMIC OBSERVATIONS

Geodynamic Institute collects and evaluates macroseismic observation for more than 100 years. The observation is published in the monthly bulletins and in such a way is distributed to the scientific community, comprising the primary material for research studies. During the last years, within the frame of various research projects, Geodynamic Institute updated the whole procedure of macroseismic observation management. More specifically:

The questionnaire has been improved, including more detailed description of damage in order for the user to be helped answering. These changes did not affect the grading according to the intensity scale used till now (Modified Mercalli, MM). Each answer has its own code number, so the answers can be introduced to statistical analysis. The questionnaire is included in the G.I. web site, in order for anybody to answer if and how he felt the earthquake and if and what damage he observed.

The questionnaire-mailing-procedure, as well as the management of the answers, are handled by a database of macroseismic observations, originally developed under MsAccess environment. The database includes three main tables with one-to-many relationships: Table SITE includes the administrative information of the municipalities and communities gathered after “Kapodistrias” law for local authorities and the recent inventory of the Hellenic Statistical Service, like the name of the municipality, the prefecture, the population, the mail and electronic address, phone numbers, as well as coordinates and the dominant surface geology (IGME maps, 1:50000). Each record has its own unique code number. The spatial distribution of the sites to which the questionnaire has been sent after the strong earthquake (according to criteria concerning the spatial distribution, the population and the epicentral distance) can be checked. Table QUAKE includes the seismic parameters of the earthquakes within the area 33^{o}N - 42^{o}N and 19^{o}E - 29^{o}E having a magnitude $M_s \geq 5.5$ occurred during the period 1900 – today. The table is updated continuously with the new strong earthquakes. Each record has its own code number, while a column is dedicated to a notification about the existence or not of macroseismic observation. Table EFFECT includes the macroseismic intensity observed at each site for the different earthquakes. Other data included here are the epicentral and hypocentral distance and the azimuth (the angle between the site-epicenter line and the line of north. Each observation is characterized by the combination of the site code and the earthquake code. An auxiliary table named INFO includes photographs, descriptions, references about the damage and the earthquakes.

The fore mentioned database has been incorporated to SEISMO-SURFER with small modification of the field names and some additional informative columns. For example the addition of tables FLINN and FLINN_RECTANGLE, which are related to the tables SITE and QUAKE, assists on the geographical representation of the affected sites using the Flinn & Engdahl geographical terminology (Young *et al*., 1996), while the addition of table COUNTRY will be used in the future to import macroseismic observation from neighboring countries (Greek earthquakes which might affect sites of Turkey, Bulgaria, FYROM or Albania) or even from other countries. Moreover table FAULTS includes details about the focal mechanism of the earthquakes (characterization of the seismogenic fault, the strike, the slip and the rake of the planes etc.), as well as the name of the seismogenic fault (ex. Servia fault), all these taken from bibliography (ex. Kiratzi and Louvari, 2003), which is also included. The Entity-Relationship diagram of figure 3 illustrates the entities included in SEISMO-SURFER and their relationships, on which the tables of Oracle relational database are based.

The integration of the macroseismic information in the SEISMO-SURFER makes the tool to be more complete and gives the advantage to the user to extract new knowledge by using its techniques taking into account greater variety of data, connected with more complicated relationships. For example, a question like “find the closest-to-a-point sites having an intensity greater than VIII and give the results together with the focal mechanism of the earthquakes, the distance and the azimuth between the sites and the epicenters, as well as the local geology” can be applied in order for the user to conclude to any possible relationship.

5 PROSPECTS AND CONCLUSIONS

In the present work the architecture and the operation of the SEISMO-SURFER is described with emphasis in the collection, management and analysis of the macroseismic data. Indicative results of this analysis are also mentioned by using knowledge-extraction techniques.

Taking into account the continuous seismological -and macroseismic – data collection by the Geodynamic Institute, as well as the comparison between the SEISMO-SURFER and other relative tools (Han *et al.,* 1997; Andrienko and Andrienko, 1999; Kretschmer and Roccatagliata, 2000) the future steps for the SEISMO-SURFER development include four directions. More specifically:

Enrichment of the database with new records and new characteristics: The enrichment of the database will include new records coming from its temporal extension to the past. This could be realized by the addition of historical catalogues (Papazachos & Papazachou, 2003). Moreover, new characteristics of the existing records will be added, like the detailed geology for the sites of macroseismic observations, or the peak ground acceleration for specific sites, the characteristics of the seismogenic fault, the seismic hazard zones and thus the connection of the information with the seismic resistant code, bibliographic references for strong earthquakes etc.

Data warehouse functionality with aggregate information: The reason for the use of warehouses instead of databases is the large volume of the seismological data, as well as the complexity of the difference queries that the user could submit. The system is unable to give answer in case of a complex query submitted to a large database. Further more, the user should have a good knowledge of the specific programming language for query formulation (SQL usually), and the functionality of the database is overloaded. In any case, it is possible to develop both a database and a data warehouse due to the differences between them: The databases are mainly used for the continuous dealing processing, while the data warehouses are used for the continuous analytical processing and the information extraction. The data within the databases are dynamic, current and detailed, while within the data warehouses are static, historical and concentrated. The modifications of the database data are continuous, while the data warehouses are subjected to periodical updates.

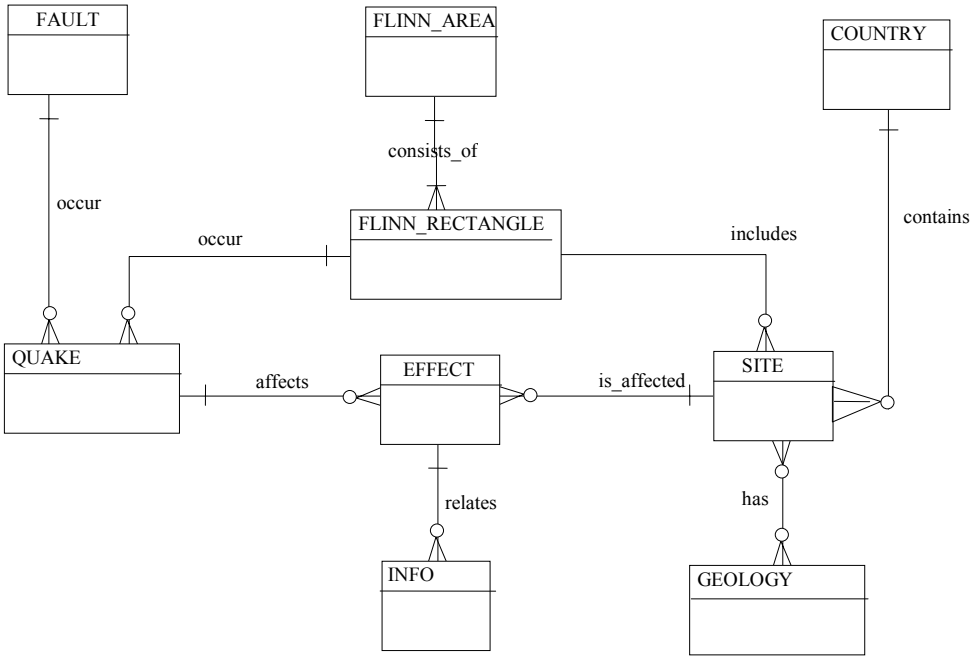

Figure 3. Diagram showing the entities included in the SEISMO-SURFER database and their relationships.

More data mining operations via additional algorithms: As already mentioned, it is only the k-means algorithm that is incorporated in the system for the moment. Nevertheless, other data mining algorithms supported by Oracle are for classification (Adaptive Bayes Network for predictive infor-

mation and Naïve Bayes), for clustering (O-cluster based on grid while k-means is based on distance), for attribute importance (predictive variance for the identification of the most effective characteristics on a parameter under investigation), for association rules (A-priori using frequent itemsets).

Various modifications – web edition: Various improvements are planned concerning the visualization part of the tool covering the possible new relations, various queries should be added or modified in order for additional parameterization of the information included, the map navigation will be improved, while the user could have the possibility to be supported by other external sources for data mining. Finally, a subset of the tool is planned to be publicly available over the Internet as a web-based portal of information and knowledge about earthquakes.

REFERENCES

Andrienko, G., Andrienko N., 1999: *Knowledge-based visualization to support spatial data mining*. Proceedings of the 3rd Symposium on Intelligent Data Analysis, Amsterdam, the Netherlands, 1999.

Comninakis, P.E. and Papazachos, B.C. (1986). A catalogue of earthquakes in Greece and the surrounding area for the period 1901 – 1985. *Univ. Thessaloniki, Geophys. Lab., publ. No. 1., 167 p.*

GI-NOA: *Earthquake Catalog*. Available at http://www.gein.noa.gr/services/cat.html.

Han, J., Koperski K., Stefanovic N., 1997: *GeoMiner: a system prototype for spatial data mining*. Proceedings of ACM SIGMOD International Conference on Management of Data, Tucson, AZ, USA, 1997.

Kiratzi, A. and Louvari, E. (2003). Focal mechanisms of shallow earthquakes in the Aegean Sea and the surrounding lands determined by waveform modeling: a new database. *Journ Geodynamics, 36, 251-274.*

Kretschmer, U., Roccatagliata E., 2000: *CommonGIS: a European project for an easy access to geo-data*. Proceedings of the 2nd European GIS Education Seminar, EUGISES, Budapest, Hungary, 2000.

NEIC-USGS: *Earthquake Search*. Available at http://neic.usgs.gov/neis/epic/epic_global.html.

Papazachos, B. & Papazachou, K. 2003. Earthquakes of Greece. 3rd edition, Ziti editions, Thessaloniki (in Greek).

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, Manolopoulos Y., Evripidou S & Kakas A. eds, Lecture Notes in Computer Science, No 2563, Springer – Verlag, Berlin.

Young, J.B., Presgrave, B.W., Aichele, H., Wiens, D.A. and Flinn, E.A., 1996. The Flinn-Engdahl Regionaligation Scheme: The 1995 Revision. Physics of the Earth and Planetary Interiors, 96, 223-297.