Geophysical Research Abstracts,
Vol. 11, EGU2009-11310, 2009
EGU General Assembly 2009

Mining and Integration of Environmental Data

V. Tran, L. Hluchy, O. Habala, and M. Ciglan
Institute of Informatics, SAS, Department of Parallel and Distributed Computing, Bratislava, Slovakia (viet.ui@savba.sk)

The project ADMIRE (Advanced Data Mining and Integration Research for Europe) is a 7th FP EU ICT project aims to deliver a consistent and easy-to-use technology for extracting information and knowledge. The project is motivated by the difficulty of extracting meaningful information by data mining combinations of data from multiple heterogeneous and distributed resources. It will also provide an abstract view of data mining and integration, which will give users and developers the power to cope with complexity and heterogeneity of services, data and processes.

The data sets describing phenomena from domains like business, society, and environment often contain spatial and temporal dimensions. Integration of spatio-temporal data from different sources is a challenging task due to those dimensions. Different spatio-temporal data sets contain data at different resolutions (e.g. size of the spatial grid) and frequencies. This heterogeneity is the principal challenge of geo-spatial and temporal data sets integration – the integrated data set should hold homogeneous data of the same resolution and frequency.

Thus, to integrate heterogeneous spatio-temporal data from distinct source, transformation of one or more data sets is necessary. Following transformation operation are required:
• transformation to common spatial and temporal representation – (e.g. transformation to common coordinate system),
• spatial and/or temporal aggregation – data from detailed data source are aggregated to match the resolution of other resources involved in the integration process,
• spatial and/or temporal record decomposition – records from source with lower resolution data are decomposed to match the granularity of the other data source. This operation decreases data quality (e.g. transformation of data from 50km grid to 10 km grid) – data from lower resolution data set in the integrated schema are imprecise, but it allows us to preserve higher resolution data.

We can decompose the spatio-temporal data integration to following phases:
• pre-integration data processing – different data set can be physically stored in different formats (e.g. relational databases, text files); it might be necessary to pre-process the data sets to be integrated,
• identification of transformation operations necessary to integrate data in spatio-temporal dimensions,
• identification of transformation operations to be performed on non-spatio-temporal attributes and
• output data schema and set generation – given prepared data and the set of transformation, operations, the final integrated schema is produces.

Spatio-temporal dimension brings its specifics also to the problem of mining spatio-temporal data sets. Spatio-temporal relationships exist among records in (s-t) data sets and those relationships should be considered in mining operation. This means that when analyzing a record in spatio-temporal data set, the records in its spatial and/or temporal proximity should be taken into account. In addition, the relationships discovered in spatio-temporal data can be different when mining the same data on different scales (e.g. mining the same data sets on 50 km grid with daily data vs. 10 km grid with hourly data).

To be able to do effective data mining, we first needed to gather a sufficient amount of environmental data covering similar area and time span. For this purpose we have engaged in cooperation with several organizations working in the environmental domain in Slovakia, some of which are also our partners from previous research efforts. The organizations which volunteered some of their data are the Slovak Hydro-meteorological Institute

(SHMU), the Slovak Water Enterprise (SVP), the Soil Science and Conservation Institute (VUPOP), and the Institute of Hydrology of the Slovak Academy of Sciences (UHSAV). We have prepared scenarios from general meteorology, as well as specialized in hydrology and soil protection.