THE ROLE OF DATA MINING TECHNIQUES IN THE INTEROPERABILITY OF GIIDA

Vito Felice Uricchio [a], Stefania D'Arpa [a], Emanuele Barca [a] Gianni Tartari [b]

[a] CNR, Water Research Institute V.le De Blasio, 5 - 70123, Bari Italy - (vito.uricchio, stefania.darpa, emanuele Barca)@ba.irsa.cnr.it

[b] CNR, Water Research Institute Via del mulino, 19 - 20047, Brugherio (MB) Italy – tartari@irsa.cnr.it

KEY WORDS: Data Mining Algorithms, Spatial data infrastructures, Interoperability, Metadata.

ABSTRACT:
Spatial Data Infrastructures (SDI) is a standard aimed to guarantee a correct structured input for accessing to information, while Data Mining Algorithms (DMA) are a powerful tool to approaching the analysis of inhomogeneous data.
The aim of the GIIDA project is the definition and improvement of a multidisciplinary shared infrastructure for managing, processing and analyzing environmental data through the implementation of a SDI complying with the international specifications. This infrastructure is designed to host the most up-to-date methodologies for exploratory data analysis such as DMA.
In concrete terms, GIIDA points to integrate the wide resources of existing data sets in a single "virtual database", making them available to all users in a easy, quick and targeted way. The GIIDA's multidisciplinary approach in different thematic environmental areas (atmosphere, oceanography, freshwaters, soil etc.) represents an interesting case study for the application of DMA.
DMA is a tool that allows one to discover previously unknown, valid patterns and relationships in large and inhomogeneous data sets, taking advantage from data integration and interoperability.
The management of complex environmental topics like risk assessment, environmental strategic evaluation, sustainable management of natural and faunal resources, can take benefit from the interaction between SDI and data mining techniques.
Within GIIDA project, DMA can be an effective way of partially compensating for the lack of metadata, through its capability of extracting valid non trivial relationships.
The coupled abilities of GIIDA infrastructure and DMA add to the final structure of GIIDA's architecture all the essential aspects of a complex environmental management information system.

1. INTRODUCTION

GIIDA (Gestione Integrata e Interoperativa dei Dati Ambientali) is a inter-departmental project of the Italian National Research Council (CNR). The project is an initiative of the Earth and Environment Department (Dipartimento Terra e Ambiente) of the CNR (Nativi, 2009). The idea of the project was born in view of the problems regarding the availability, quality, organization, accessibility and sharing of environmental information.
To solve these problems and to respond to the need to strengthen the presence of the CNR in national and international context, the project GIIDA pursues the creation of an infrastructure for shared environmental information in order to unify and integrate the wide resources of environmental information, existing on national territory, in a single "virtual database" and make it available to all users in an easy, quick and targeted way.
The project, therefore, aims to establish an environmental Spatial Data Infrastructure (SDI) running on tens of different CNR databases, by creating a multidisciplinary cyber-infrastructure for managing, processing and evaluating earth and environmental data, where spatial data, metadata, users, tools, and services are interactively connected within a ruled framework, in order to allow one using data efficiently and flexibly.
To achieve this "System of Systems", GIIDA applies a mediation framework to federate thematic sub-systems serving different Societal Benefit Areas.
These sub-systems (i.e. infrastructures) are called GIIDA thematic systems and refer to the following thematic areas:
• Biodiversity;
• Climate Changes;
• Air Quality;
• Soil and Water Quality;
• Risks;
• Infrastructures for Research and Public Administrations;
• Sea and Marine resources.
Figure 1 summarizes the organizational structure of GIIDA project.
Different working groups of experts have been created for each thematic area with the aim of developing activities, respecting timetables and gaining expected achievements.
Main activities of the above cited working groups consist in developing:

1) a specific Web Portal;
2) a thematic catalogue service;
3) a thematic thesaurus service;
4) a thematic Wiki;

defining:

5) standard access and visualization facilities for the thematic resources such as datasets;

and providing

6) models, and data processing facilities;
7) significant operative scenarios to be tested.

Aim of this work is to shortly present one of the nearing completion data access facilities developed by the Soil and Water Quality working group of GIIDA, based on Knowledge Discovery in Databases (KDD) techniques. In particular, the proposed methodology calculates the linear coefficient of correlation and the associated scatterplot in the context of a distributed database. This service has been implemented taking advantage from the peculiar design of the GIIDA infrastructure strongly based on the metadata management.
The KDD techniques, as the proposed is, belong to wider field of the Data Mining, which is a discipline providing some techniques aimed to the improvement of decisional processes

through data analysis. therefore, this work can be viewed as one of the first attempts of creating a data mining service within GIIDA project.

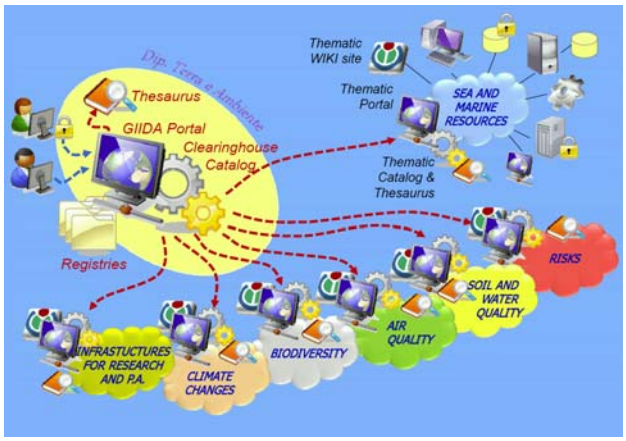

Figure 1: Organizational structure of GIIDA project.

2. THE ARCHITECTURE FOR INTEROPERABILITY OF GIIDA

GIIDA infrastructure architecture has been designed and developed according with international specifications, pursuing global interoperability of environmental information.
The interoperability is a fundamental requirement of SDI and it is referred to the possibility of combining spatial data sets, and developing system interaction services capable of avoiding repetitive manual operations, while maintaining consistent results and enhancing added value of datasets and related services. (Inspire,2007).
The systems interoperability is aimed in order to facilitate interactions between different and non-homogeneous systems, both because of software and hardware.
Finally, in order to achieve the global interoperability of environmental information, the architecture of the GIIDA's distributed infrastructure, refers to architectural solutions developing within geomatics standardization initiatives (eg. ISO, OASIS, INSPIRE, OGC, GEOSS).

3. METADATA AND INTEROPERABILITY

Adopting metadata standards is a key aspect in order to achieve a full interoperability between heterogeneous information systems.
Metadata provides an efficient and effective representation oB data checking at the same time, the exchange of information between entities that cooperate to achieve a specific goal.
By definition, indeed, the term "metadata" refers to all information related to data, about data and around data and its structured information, describe, explain and place an information resource in the context from which it came (Day, 2001).
According to their specific function metadata can be:

- Descriptive: when they allow to identifying resources, facilitating their submission and retrieval;
- Administrative: when they allow resources organization, management, localization and provide resources statistics;
- Structural: when they allow resources storage.

If implemented according to standard formats, metadata allow the interoperability and integration of similar resources; in technical terms, the establishment of geographically distributed databases.
Data exchanging among systems with different hardware and/or software characteristics, often involves loss of contents and functionality. The solution is to adopt predefined metadata schemes, shared transfer protocols and crosswalks (mapping) between different schemes of metadata.
Considering shared database, interoperability is ensured by the systematic collection (harvesting) and centralized indexing of metadata, through appropriate interrogation and data exchange protocols. This allows the sites to retrieve or "harvesting" the metadata from various resources, and use it to offer services such as indexing or link services. Such a service may allow users to access information from a large number of storage sites through a single central node.
The data providers export metadata in a common format, often encoded in XML. The service providers (harvesters, aggregators, caches, proxies, gateways) collect, process and index the metadata through the default protocol. From the metadata repository is also possible, in the opposite direction, to go back to datasets that share similar characteristics, through a quick search in centralized and compact repositories avoiding direct queries in each of the peripheral database components.
Uniform resource identifier guarantees the unique identification of the objects contained in the original database through their metadata and the access to each item.

4. GIIDA INFRASTRUCTURE AND DATA MINING TECHNOLOGIES

The architectural design of GIIDA contains a specific computational view aimed to host development environments for data processing and information extraction which include the generation of value-added information and knowledge discovery. In this context Data Mining Technologies (DMT) take a considerable importance.
Data mining is a comprehensive term which indicates a set of automatic seeking techniques in massive datasets and methods for discovering unknown and non-trivial relationship among them and summarizing them in understandable and useful ways. (Hand et al., 2001).
For this reasons, DMT is often associated with the process of Knowledge Discovery (KD). DTM, in fact, considered as an autonomous discipline, can be placed at the intersection among computer sciences, statistics and artificial intelligence. The use of such techniques is taking great advantage from recent trend of creating large strategic database (data warehouse) and extracting, clean and standardized information from them, providing information systems that allow to search for potential connections between previously separate and not comparable information. The size of cases (records) and the large amount of observed characteristics (variables) contained in these databases require specific techniques to be applied able to take into account the computational challenge related to handling such large datasets and that this datasets are often not statistically representative sample.
In consideration of these aspects statistical techniques to be applied to large datasets, need to be designed in order to avoid the risk of generating doubtful or unreliable results.

5. THE DATA MINING SERVICE

This work demonstrates how DMT, became more efficient when applied to a "System of Systems" like GIIDA where datasets have been integrated and made interoperable.
In the following sections a data mining service designed by IRSA-CNR to be hosted within the GIIDA system is described.
The service will benefit from metadata standards adopted by the SDI, complying with national and European directives on the digitization of information resources, and will leverage the interoperability between the databases involved in the project.
The algorithm implemented in the data mining service uses a well-known statistical model, the correlation, and makes it suitable to handle large datasets within the GIIDA infrastructure.
Carrying out a correlation analysis means to compare two distinct datasets, ordered by pairs, to measure their degree of association. The comparison is expressed by means of a single value, called Linear Correlation Index (LCI) which can provide useful suggestions, , under particular hypothesis, about possible cause-effect relationships between the processes generating the compared datasets
The designed service provides together with the Pearson linear correlation index which is a quantitative measure of the explored correlation, also a graphical representation (scatterplot) of the correlation between couples of datasets The scatterplot is a graphic representation of data, providing a very similar information as correlation does, but in qualitative (visual) form.
The scatterplot contains in a defined Cartesian coordinate system all the couples (x, y) coming from the two datasets.
The Pearson linear correlation index (Neter et al, 1996) between two random variables X and Y is defined as the ratio between their covariance and the product of standard deviations of both variables.
The index formula can be expressed as follows:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad (1)$$

where σ_{xy} = covariance between the two sets of data;

σ_x, σ_y = standard deviations associated with the two series.

The correlation index is an adimensional one because of its own structure, so it is possible to compare datasets with different measurement units with falling into incoherent results. Moreover, its multiplicative structure makes the index a symmetric one, therefore the following relationship is valid:

$$\rho_{yx} = \rho_{xy} \qquad (2)$$

That is to say, exchanging the order of the two datasets compared the final result stays the same. Finally, the correlation coefficient values, because of the relationship between the covariance and the two standard deviations, can vary only within the [-1, 1] interval.
The compared data series have to be interpreted as uncorrelated if $\rho = 0$, the correlation is maximum if $\rho = 1$ or $\rho = -1$, in the first case the correlation is called direct in the second one it is called indirect. When the correlation is direct, the ordered pairs are characterized either by large values or small values, in the indirect case vice versa happens.
In order to apply this methodology, the datasets to be compared must share some similar property; they must share the same size, at least. Nevertheless, this last property, does not guarantee, in general, that two different datasets are comparable in terms of correlation index. Indeed, forcing the comparison between them can lead to the phenomenon known as data dredging (data snooping or data fishing). In other words, it may lead to an unsound correlation. In order to avoid the risk of falling in false correlation, it is often necessary to further restrict the datasets to be compared applying, for instance, constraints related to time or space sharing.
Consequently, at a first stage, the DM service for GIIDA searches within the metadata in order to find the datasets sharing the space/time constraints described above. Once datasets have been found, they are uploaded from the peripheral databases in a suitable format for the following step of processing.
The DM service provides two different tools for visualizing data processing results: - a correlation table consisting in symmetric matrix of correlation indices, between all the datasets couples; - a table containing the corresponding scatterplots.

In the following are shown two possible output of the service.

	El.C. 20°	***Cl***	***SO4***
El.C. 20°	1.00	1.00	0.96
Cl	1.00	1.00	0.95
SO4	0.96	0.95	1.00

Table I: A correlation matrix

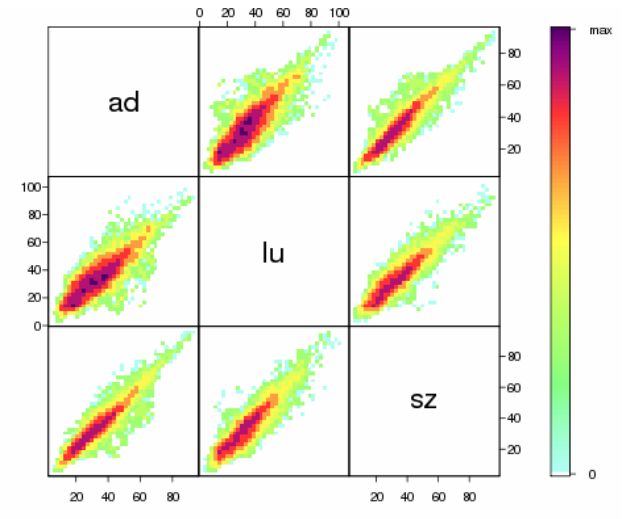

Figure 2: A scatterplot table.

6. CONCLUSION

Creating a data warehouse, merging the contribution of different data banks, is often considered as a way to share datasets with a wider user's basin than they who were able to

access them earlier. But if so, it does not justify the effort and the economic cost associated to accomplishment of this task. The creation of a distributed database to be really effective must provide a set of services able to extract non-trivial knowledge from it. These set of services will provide, on one hand, a benchmark to measure the effectiveness of the distributed database and, on the other hand, a tool to the researchers to carry out their works in a speeder way. The philosophy of GIIDA is just based on this paradigm: providing added value information and not only data to the users. The suitable structure of metadata designed for the project allow to provide services with the above described characteristics. In this background, has to be framed the present work that contains a first example of application of KD techniques to the context of GIIDA project. The implementation strategy takes advantage from the designed infrastructure for metadata, in fact, it guarantees to complete interoperability between data coming from different data providers, so allowing the comparison of datasets coming from different origins. In particular, the applied methodology assesses the linear correlation between couples of datasets and sketches its graphical counterpart: the scatterplot. As it is well known, not all the datasets can be compared in this way, so the methodology is articulated in two stages: the first one looks for comparable datasets and the second one carries out the statistical calculations. The first stage is evidently the most time consuming if applied on the real datasets, the proposed methodology, instead, processing the metadata that are a smaller set if compared to the whole dataset, shrink significantly the processing time.

7. REFERENCES

References from Books:

Hand, D., Mannila, H., Smyth, P., 2001. *Principles of Data Mining*. The MIT Press, Cambridge Massachusetts London England.

Neter, J., Kutner, M. H., Nachtsheim, C. J., 1996. *Applied Linear Statistical Models* (4th edition). Chicago, IL: Richard D. Irwin, Inc.

References from Other Literature:

Day, M. (2001). *Metadata for digital preservation: a review of recent developments*. In: P. Constantopoulos and I. T. Sølvberg, (eds.), Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001, Proceedings, Lecture Notes in Computer Science, 2163, Berlin: Springer-Verlag, 2001, pp. 161-172. ISBN 3-540-42537-3

Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE).

References from websites:

Nativi S., 2009. The GIIDA (Management of the CNR Environmental Data for Interoperability) project. Geophysical Research Abstract, Volume 11, EGU2009-3425. http://meetingorganizer.copernicus.org/EGU2009/EGU2009-3425.pdf (accessed 19 nov 2009)