Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

An integrated platform for spatial data mining and interactive visual analysis

M. May & A Savinov
Fraunhofer Institute for Autonomous Intelligent Systems, Knowledge Discovery Team, Germany

Abstract

Data Mining and Geographic Information Systems (GIS) have existed so far as separate technologies. The overall objective of the SPIN!-project is to develop a web-based spatial data mining system by integrating state of the art (GIS) and data mining functionality in a closely coupled open and extensible system architecture. The general architecture for the SPIN! spatial data mining system is described.

1 Introduction

It is estimated that 80% of data are geo-referenced. Yet most data mining systems ignore the spatial dimension. On the other hand, contemporary Geographic Information Systems (GIS) have only very basic analysis functionality. This is why data mining combined with GIS offers great potential benefits for solving the problem of spatial data analysis. Although there exist a few attempts to integrate data mining and GIS [4,11] this is a largely unexplored research area.

This paper describes an open, extensible architecture for spatial data mining. It integrates Geographic Information System for interactive visual data exploration and Data Mining functionality specially adapted for spatial data. The architecture of the system will be described, design choices will be discussed. The system is built on the Enterprise Java Bean Architecture (EJB). EJB is a server-side component architecture based on the Java 2 platform. It cleanly separates the "business logic" (the analysis tools, in our case) from server issues,

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

shielding the method developers from many technicalities involved in client-server programming. This choice allows to meet the requirements often found in business applications, e.g. security, scalability, platform independence, in a principled manner. The system is tightly integrated with a relational database and can serve as data access and transformation tool for spatial and non-spatial data. Analysis tools can be integrated either as stand-alone modules or, more tightly, by distributing the analysis functionality between the database and the core algorithm.

The final system integrates several data mining methods adapted to the analysis of spatial data, e.g., multi-relational subgroup discovery and spatial cluster analysis, and combines them with thematic mapping functionality for visual data exploration, thus offering an integrated environment for spatial data analysis.

2 Combining Data Mining and GIS

Geographic Information Systems (GIS) are widely used for analyzing and visualizing geo-referenced data. In the last few years, a new generation of Geographic Information Systems has emerged that extends the interactivity of dynamically generated maps, greatly enhancing visual exploratory data analysis ([1], [3], [6], [15]). While being an exciting development for automating cartography, these systems have limited capabilities to visualize attribute interaction on a map having more than a few dimensions. Hence, complex multi-variate dependencies are easily overlooked.

Searching for multi-variate dependencies is where *data mining* promises great benefits. Data mining is the partially automated search for hidden patterns in typically large and multi-dimensional databases. It draws on results in machine learning, statistics and database theory. Some data mining methods, such as *k*-nearest neighbor, are extensions of statistical techniques known for a long time. Others, especially from the area of machine learning and inductive logic programming (ILP), are essentially new (cf. [10]). These techniques have been packaged in *data mining platforms*, which are software environments providing support for the application of one or more data-mining algorithms.

So far Data Mining and Geographic Information Systems (GIS) have existed as two separate technologies, each with its own methods, traditions and approaches to visualization and data analysis. Recently, the task of integrating these two technologies has become highly actual especially as various public and private sector organizations possessing huge databases with thematic and geographically referenced data began to realize the huge potential of information hidden there. Among those organizations are

- statistical offices wanting to analyze or disseminate geo-referenced statistical data,
- public health services searching for explanations of disease clusters,
- environmental agencies assessing the impact of changing land use patterns on climate change,

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

- geo-marketing companies doing customer segmentation based on spatial location.

As a response to this demand a prototype has been developed [2] which demonstrates the potential of combining data mining and GIS. This initial prototype encouraged the development of the SPIN! [5,14] system the overall objective of which consists in developing a web-based spatial data mining platform by integrating state of the art Geographic Information System (GIS) and data mining functionality in a closely coupled open and extensible system architecture. The new generation SPIN! system pays special attention to such features as scalability, security, multi-user access, robustness, platform independence and adherence to standards. In this paper, we describe the general architecture of the SPIN! data mining platform.

What benefits does data mining offer for the GIS user? Data mining and geographical information systems are best seen as complementary tools for describing and analyzing data. Whereas in GIS the user guides the search and generates hypotheses, data mining partially delegates this task to the computer, pre-selecting and presenting to the analyst only those patterns deemed most interesting (according to some measure of quality). Whereas GIS relies on visualization in geographical space, data-mining searches for patterns in multi-dimensional abstract space. Both techniques are essentially exploratory, leaving the final decision of whether a hypothesis is an important new finding (a "nugget" in data mining language) or just an artifact to the analyst.

How are spatial data handled in usual data mining systems? Although many data-mining applications deal at least implicitly with spatial data they essentially ignore the spatial dimension of the data, treating them as non-spatial. This has ramifications both for the analysis of data and for their visualization. First, one of the basic tasks of exploratory data analysis is to present the salient features of a data set in a format understandable to humans. It is well known that visualization in geographical space is much easier to understand than visualization in abstract space. Secondly, results of a data mining analysis may be sub-optimal or even be distorted if unique features of spatial data, such as spatial autocorrelation ([7]), are ignored.

In sum, convergence of GIS and data mining in an Internet enabled spatial data mining system is a logical progression for spatial data analysis technology. Related work in this direction has been done by Koperski and Han, Ester et al. [4,12].

3 SPIN!: the elements

To describe the functionality of the SPIN!-system, it is useful to distinguish several levels of functionality.

Level 1: Data access and management. The basis functionality provides *data access* to heterogeneous data sources, *data transformation* capabilities, and facilities for *organizing* and *documenting* analysis tasks.

Level 2: Interactive thematic mapping for visualizing statistical data. For visual exploratory spatial analysis the Descartes module for interactive

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

manipulation of statistical maps is used ([1]). It supports basic GIS operations such as zooming, panning, querying features and changing visual appearance. Yet its real strength lies in its capabilities for interactive visual exploration of statistical data. Descartes automates map design by incorporating the knowledge of thematic cartography in the form of generic, domain-independent rules, taking into account data characteristics and relations among data components or attributes. The automation of map generation releases the user from the necessity of thinking about how to present the data and from the routine work of map building; instead she can concentrate on the analysis of her data. Among Descartes features are linked displays, interactive cross-classification, box-plots and a module for temporal visualization.

Level 3: Spatial cluster detection. Descartes can be used for interactive, visual identification of spatial clusters. Yet the SPIN!-system also contains modules for performing this search automatically. The objective of the *Geographical Analysis Machine* GAM [15] is to look for local spatial clusters without knowing in advance where to look. GAM works by examining a large number of overlapping circles of varying sizes that completely cover a region of interest, retaining cycles with a statistically significant deviation in distribution.

Level 4: Explaining clusters and spatial phenomena. Assume we have found a spatial cluster or interesting classification, using either the interactive approach of Descartes or the automated search of GAM. *What attributes are associated with a cluster that could potentially explain it*? To answer this question, spatial data mining methods are applied. The key to spatial data mining is to make proper use of spatial information inherent in the data by extending the representational capabilities of data mining algorithms. While traditional attribute-value based learning methods have difficulties in expressing topological features such as `close_to`, `adjacent_to` etc. in a natural and general way, they can be easily expressed in first-order-logic. This makes *inductive logic programming* (ILP), which uses a first-order representation, a natural and promising approach to many forms of spatial data mining. In the SPIN! project we investigate spatial association rules and subgroup discovery ([8], [13], [16]).

The way data mining results are presented to the user is crucial for their appropriate interpretation. We use a combination of cartographic and non-cartographic displays linked together through simultaneous dynamic highlighting of the corresponding parts (Fig. 1).

4 n-tier EJB-based architecture

The general SPIN! architecture is shown in Fig. 2. It is a *n*-tier Client/Server-architecture based on *Enterprise Java Beans* for the server side components. A major advantage of using Enterprise Java Beans is that tasks as controlling and maintaining user access rights, handling multi-user access, pooling of connections, caching, handling persistency and transaction management are delegated to the *EJB container*. The architecture has the following major subsystems: *client*, *application server* with one or more EJB containers, one or more *database servers* and optionally *compute servers*.

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Client. The *client* is a GUI Java application or applet. It always creates one server side representative in the form of session bean the methods of which are accessed either directly through the corresponding remote reference (Java RMI or CORBA IIOP protocol) or indirectly by means of servlets (HTTP protocol). The client session bean executes various server side tasks on behalf of the client. In particular, it may load/save workspace objects from/in its persistent state.

The client is based on component connectivity conception, which is implemented in Java as connectivity library (CoCon). The idea is that the workspace consists of components each of which is considered a storage for a set of parameters and pieces of functionality (e.g., algorithms). The system functionality is determined by a set of available components.

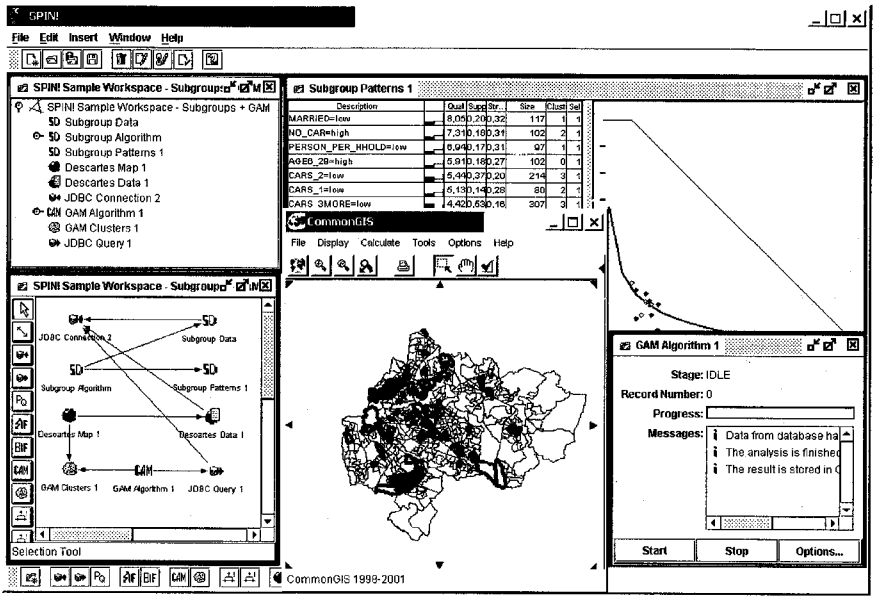

Figure 1: SPIN! client. The workspace consists of interconnected components such as database connections, database queries, data mining algorithms, analysis results and spatial object visualizers.

The components can be connected via different types of connections. In the SPIN! system three types of connections have been used: *hierarchical* (parent/child or vertical) connections, *user* (or horizontal) connections, and *view* (or visualization) connections. Hierarchical connections serve to implement the hierarchical workspace structure and are added or removed automatically by the system while adding/removing components. The view connections specify current the visualizer for each workspace component and are also created and deleted automatically as windows for components are opened/closed. The user

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

connections are intended to connect components, which need to cooperate to perform some task and are created by the user according to the workspace purpose. For example, the typical scenario is to connect a data mining algorithm with a data description (query) and a result component. What is even more important is that components know how they can be correctly connected so that only valid configurations can be composed by the user. For example, it is not possible to connect an algorithm to an inappropriate visualizer.

In the client there are two views for editing workspaces: the tree view and the graph view. In the tree view components from the system repository can be added into the workspace (fig. 1, left top). User connections can be established in the Connection Editor dialog. A more user friendly way of editing workspaces is through a Clementine-style workspace graph view, which shows both components and their user connections (fig.1, left bottom). In this view components can be added by selecting them from the tool bar and connecting them by drawing arrows between graph nodes. It is also very important that components can be arranged within views into visually expressive diagrams.

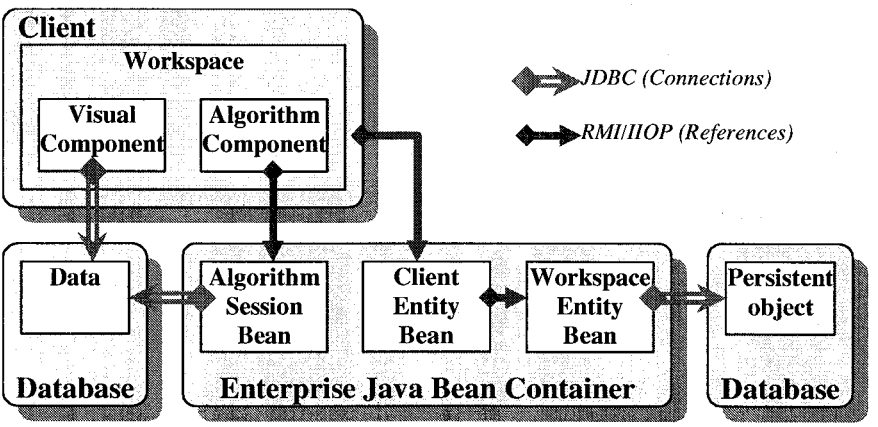

Figure 2: SPIN! platform architecture. Main components are a Java-based client, an Enterprise Java Beans Container and one or more databases serving spatial and non-spatial data.

Application server. The *application server* is an Enterprise Java Bean container. It manages the client workspace, analysis tasks, data access and persistency. There may be more than one simultaneously running container on one or more servers so that, e.g., different algorithms and other tasks can be executed on different computers under different restrictions. The SPIN! system uses an EJB container for making workspaces persistent in the database and for remote computations. For the first task the client creates a special session bean, which is responsible on the server side for workspace persistence and access. Particularly, if the client needs to load or save a workspace it delegates this task

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

to this session bean. The simplest strategy for storing workspace is to serialize it and save it in a record as one large object. A more sophisticated approach is to use two tables: one for workspace components and one for connections between components. Each workspace component and connection is saved into one record. To retrieve the workspace the session bean loads all component and connections objects and then reproduces the workspace run time graph structure.

The client creates one remote object for each analysis task to be run so that data is transferred directly from the database to the algorithm. After the analysis is finished its result is transferred to the client for visualization. A *connector machine*, which is a Java Virtual Machine running on the application server, is used for accessing non-Java analysis tasks. Those may run on additional compute servers.

Data storage. User data are stored in primary *data storage*, which is a relational database system (it may be the same machine as the application server). There may be one or more optional *secondary databases* for analysing data. In addition, data can be loaded from other sources – databases, ASCII files in the file system or Excel files. It is important that for remote computations in application server data is transferred directly into the remote algorithm bypassing the client. It is only a set of components (subgraph of the workspace) that is transferred between application server and client.

5 Running Data Mining algorithms

The developed architecture supposes that all algorithms are executed on compute servers. For each running algorithm a separate session bean is created which implements high-level methods for controlling its behaviour, particularly, starting/stopping the execution, getting/setting parameters, setting the data to process, and getting the result. The session bean then is responsible for the methods implementation. There are several ways how it can be done.

- A clean and very convenient but in some cases inefficient approach is using Java for implementing the complete algorithm directly within the corresponding EJB, loading all data via JDBC into the work-space.
- A second approach divides the labor between the EJB container and the relational database. We have implemented a multi-relational spatial subgroup-mining algorithm that does most of the analysis work (especially the spatial analysis) directly in the database. The EJB part retrieves summary statistics, manages hypotheses and controls the search.
- A third approach consists in implementing computationally intensive methods in native code wrapped into shared library by means of Java Native Interface (JNI).
- A fourth option is that the algorithm session bean directly calls an external executable module with a set of parameters to carry out its procession task.
- And finally other remote objects (e.g. CORBA) can be used to execute the task.

The algorithm parameters are formed in the client and transferred to the algorithm EJB as a workspace component before the execution. In particular,

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

data to be processed by the algorithm has to be specified. It is important that only a data *description* is specified and not the complete data set is transferred. In other words, the algorithm bean gets information where and how to take data and what kind of restrictions to use. Thus when the algorithm starts, the data is directly retrieved by the algorithm EJB rather than passes through the client.

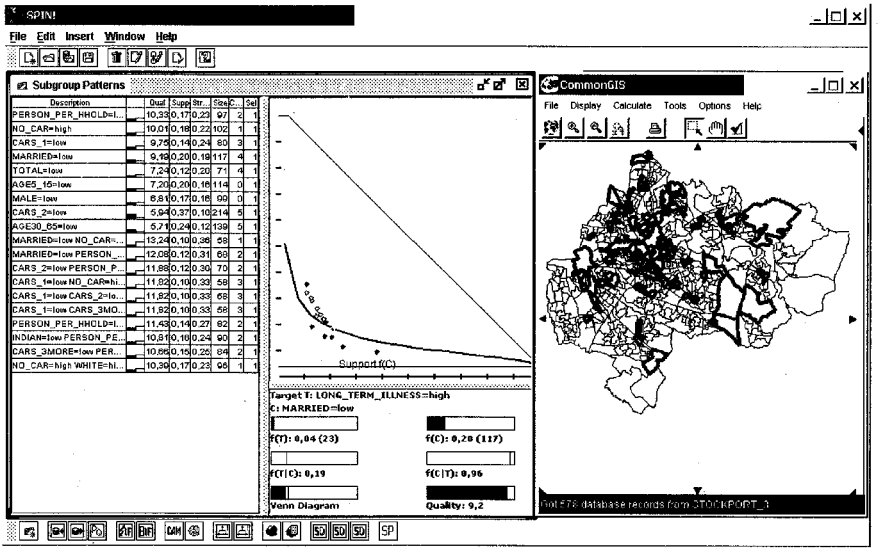

Figure 3: Spatial objects satisfying the currently selected subgroup in the left view (spatial data mining result) are highlighted on the map (GIS).

For example, assume that we need to find interesting subgroups in spatially referenced data [9]. The data is characterised by both thematic attributes, e.g., population, and spatial attributes, e.g., proximity to highway or percentage of forests in the area. The data to be analysed is specified in the corresponding component where we can choose tables, columns, ´join and restriction conditions including spatial operators supported by the underlying database system. The algorithm component is connected to the data component and the subgroup pattern component. The algorithm can be started either in local or remote computation mode. In local computation mode the analysis is carried out on a local computer. In remote mode the algorithm component creates a remote algorithm object in the EJB container as a session bean and transfers to it all necessary components such as the data description. The remote object (EJB) starts computations while its local counterpart periodically checks its state until the process is finished. During computations the remote object retrieves data, analyses it and stores the result in another component. Note that each client may start several local and remote analysis algorithms simultaneously; for each of them a separate thread is created. In local mode this thread carries out real

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

computations while in remote mode the thread looks after the remote process state. Once interesting subgroup have been discovered and stored in a component they can be visualised in a special view, which provides a list of all subgroups with all parameters as well as a two-dimensional chart where each subgroup is represented by one point according to its coverage and strength (left window in Fig. 3).

Alternatively, the data analysed by subgroup discovery data mining algorithm can be viewed in a geographic information system and analysed by visual analysis methods. For this purpose we insert into the workspace a component, which describes a) the data to be visualised and b) the map component. When the map is opened it loads data from the component, which is connected to it. If we connect more data components then each of them will be loaded into one geographic layer (right window in Fig. 3). Now it is possible to highlight objects on the map while selecting subgroups in another window. The mechanism of simultaneous highlighting is based on using the same identifiers specified for both components and propagating the corresponding events through the workspace while the current selection changes.

6 Conclusion

We have described the general architecture of the SPIN! spatial data mining platform. It integrates GIS and data mining algorithms that have been adapted to spatial data. The choice of EJB technology allows us to meet requirements such as security, scalability, platform independence, in a principled manner. The system is tightly integrated with a RDBMS and can serve as data access and transformation tool for spatial and non-spatial data.

Acknowledgement: Work on this paper has been partially funded by the European Commission under IST-1999-10536-SPIN!

References

[1] Andrienko, G.; Andrienko, N.. "Interactive Maps for Visual Data Exploration", *International Journal of Geographical Information Science* 13(5), 355-374, 1999

[2] Andrienko, N., G. Andrienko, A. Savinov, and D. Wettschereck, "Descartes and Kepler for Spatial Data Mining", ERCIM News, No. 40, January 2000, 44–45.

[3] Dykes, J., "Exploring spatial data exploration with dynamic graphics", *Computers and Geosciences*, 23, 345-370, 1997

[4] Ester, M., Frommelt, A., Kriegel, H.P, Sander, J., "Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support", in *Data Minining and Knowledge Discovery, an International Journal*, 1999

[5] European IST SPIN!-project web site, http://www.ccg.leeds.ac.uk/spin/

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

[6] Gitis V., Dovgyallo A., Osher B., Gergely T., "GeoNet: an information technology for WWW on-line intelligent Geodata analysis", *Abstracts of 4th EC-GIS Workshop*, Hungary, 1998
[7] Haining, R. Spatial data analysis in the social and environmental sciences, Cambridge Univ. Press, 1991
[8] Klösgen, W., "Deviation and association patterns for subgroup mining in temporal, spatial, and textual data bases", In: Polkowski, L., Skowron, A. (eds): *Rough sets and current trends in computing*, 1-18, New York, Springer, 1998
[9] W. Klösgen, May, M. Spatial Subgroup Mining Integrated in an Object-Relational Spatial Database, PKDD 2002, Helsinki, in press 2002
[10] Klösgen, W., Zytkow, J. (eds.), Handbook of Data Mining and Knowledge Discovery. Oxford University Press, to appear 2001
[11] Koperski, K., Adhikary, J., Han, J., 1996. Spatial Data Mining, Progress and Challenges, Vancouver, Canada, Technical Report
[12] Koperski, K., Han, J. "GeoMiner: A System Prototype for Spatial Mining", Proceedings ACM-SIGMOD, Arizona, 1997
[13] Malerba, D.; Esposito, F., Lisi, F., "A logical framework for frequent pattern discovery in spatial data", Proceedngs of the 14th International FLAIRS Conference, accepted, 2001
[14] May, M.: Spatial Knowledge Discovery: The SPIN! System. Fullerton, K. (ed.) Proceedings of the 6th EC-GIS Workshop, Lyon, 28-30th June, European Commission, JRC, Ispra.
[15] Openshaw, S., Turton, I., Macgill, J. and Davy, J., "Putting the Geographical Analysis Machine on the Internet", in Gittings, B. (ed.) *Innovations in GIS 6,* Taylor and Francis, London, 1999
[16] Wrobel, S. "Scalability Issues in Inductive Logic Programming", In *Proc. 9th Int. Workshop on Algorithmic Learning Theory* (ALT-98), Berlin, Springer, 1998