

Towards a flexible system for exploratory spatio-temporal data mining and visualization

S. DI MARTINO^{a*}, F. FERRUCCI^a, M. BERTOLOTTO^b, T. KECHADI^b

^aUniversità degli Studi di Salerno – DMI, Fisciano (SA), Italy
{sdimartino, fferrucci}@unisa.it

^bUniversity College Dublin, Dublin, Ireland
{michela.bertolotto, tahar.kechadi}@ucd.ie

Many natural phenomena present intrinsic spatial and temporal characteristics. With the recent advances in data collection technologies, high resolution spatio-temporal datasets can be stored and analyzed to accurately study the behavior of such events. However, these datasets are often very large and difficult to analyze and display. Recently much attention has been dedicated to the application of innovative data-mining techniques to filter out relevant subsets of very large repositories as well as to the development of visualization tools to effectively display the corresponding results. In this paper we describe our approach to deal with very large spatio-temporal datasets. Our framework includes new techniques to efficiently support the data-mining process, address the spatial and temporal dimensions of the dataset, and visualize and interpret results. Within this framework, we have developed two complementary 3D visualization environments, one based on Google Earth and one relying on a Java3D graphical user interface. In this paper we provide an overview of the system we have developed and we highlight the challenges we are dealing with, to handle a new, wide dataset containing heterogeneous multi-dimensional information on traffic events.

Keywords: data mining; spatio-temporal data; exploratory visualization.

1. Introduction

Some research estimates that about 80% of the data stored in corporate databases integrate spatial information (Fayyad and Grinstein, 2001), leading to huge amounts of geo-referenced information that need to be analyzed and processed. These datasets are often critical for decision support, but their value depends on the ability to extract useful information for studying and understanding the phenomena governing the data source. Therefore, the need for efficient and effective techniques for analyzing spatio-temporal datasets has recently emerged as a research priority (Bédard et al, 2001): spatio-temporal Data Mining aims at addressing these needs. It encompasses a set of exploratory, computational and interactive approaches for analyzing very large spatial and spatio-temporal datasets. Numerous research projects on spatial data mining have been conducted in the last two decades (a comprehensive review is provided by Andrienko et al., 2003). Several open issues have been identified, ranging from the definition of mining techniques capable of dealing with spatial-temporal information, to the development of effective methods for interpreting and visualizing the final results. In particular, visualization techniques are widely

recognized to be powerful in this domain (Andrienko et al., 2003), (Andrienko et al., 2005), (Johnston, 2001), since they take advantage of human abilities to perceive visual patterns and to interpret them (Andrienko et al., 2003), (Kopanakis and Theodoulidis, 2003), (Costabile and Malerba, 2003). However, it is recognized that spatial visualization features provided in the existing geographical applications are not adequate for decision support when used alone. Hence, alternative solutions have to be defined (Bédard et al, 2001), to dynamically and interactively obtain different spatial and temporal views, and to interact in different ways with the results produced during the data mining process. The problems of how to visualize the spatio-temporal multidimensional dataset (Bédard et al, 1997) and how to define effective visual interfaces for viewing and manipulating the geometrical components of the spatial data (Shneiderman, 2002) are some of the challenges that still need to be tackled.

To address these issues, we developed a system for decision-centered visual analysis of spatio-temporal datasets. The aim of this system is, on one hand, to enable data-mining tools to provide some form of localization in the data being analyzed, and, on the other hand, to interactively visualize in 3D the outcome of the mining process. To achieve these goals, the system includes a data-mining engine that can integrate different data-mining algorithms (to work with specific types of datasets) and two complementary 3D visualization tools. One exploits Google Earth (Google Earth website, 2005) to render in 3D the mining outcomes over a geo-referenced satellite image, enhanced by additional informative layers. The other visualization tool exploits Java3D (Java3D Website, 2001) to provide more advanced user interaction with the mining results, by providing a set of features oriented to data mining experts. This system was successfully tested against a huge real-world dataset (Hurricane Isabel, which struck the US east coast in September 2003, see (National Hurricane Center, 2003)), as described in (Compieta et al., 2006). The corresponding dataset was about 62.5GB, containing more than 25 millions real-valued points in each time-step. The system allowed us to detect both expected and unexpected behaviors, as well as to find interesting relationships and specific patterns/characteristics about hurricane data.

By exploiting the flexibility of the proposed system architecture, we are currently working towards adapting the system to deal with a different spatio-temporal dataset, provided by the Italian highways agency and including all events that occurred in the last 3 years over a 50 km stretch of highway, characterized by critical traffic density. This wide and dynamic dataset contains heterogeneous multi-dimensional information, describing thousands of events including accidents, traffic flow, exceptional

events, meteorological phenomena, etc., partly basing on DATEX (DATEX Standard Specifications, 2005), a traffic and travel data exchange standard, designed to uniformly model interactions between inter-regional Traffic Control Centers in Europe. Our system could be very useful to discover undetected correlations between phenomena (like meteorological events, shape of the highway, specific levels of traffic flow, sun position, etc.). As a consequence, adequate improvements could be taken, aimed at reducing infrastructure hazards, and new traffic policies could be defined to reduce the number of accidents, and thus save human lives. Clearly, this different dataset is posing many challenges, as the system has to be adapted to deal not only with points, but also with linear data (e.g., the trajectories of the vehicles involved in accidents). Thus, we are defining adequate methods to mine, visualize, and exploit this heterogeneous information.

In this paper we provide an overview of the system we have developed and experimented with the meteorological application, and we highlight the challenges we are dealing with, to handle the new dataset. The remainder is structured as follows. Section 2 describes the architecture of the proposed system. Section 3 is devoted to describe the specifically suited data-mining engine, while in Section 4 the two visual environments are presented. Finally, in Section 5 we describe the current research direction.

2. System Architecture

The Data Mining process usually consists of three phases: 1) pre-processing (data preparation), 2) modeling and validation and, 3) post-processing (deployment). During the first phase, the data may need some cleaning and transformation according to some constraints imposed by some tools, algorithms, or users. The second phase consists of choosing or building a model that better reflects the application behavior. Finally, the third step consists of using the model, evaluated and validated in the second phase, to effectively study the application behavior. Usually, the model output requires some “post-processing” in order to exploit it. This step can take all the benefits of visual analytics, since interactivity and user expertise are very important in the final decision-making and data interpretation. Based on these phases, the proposed system relies on a three-tier architecture, including a data store at the back end, an application server, and two visualization components at the front end (see Figure 1). The application server runs all application programs that perform the mining tasks. The mining engine produces an output model that contains structured results. Depending on a specific adopted mining algorithm, these results may need to be manipulated in different ways.

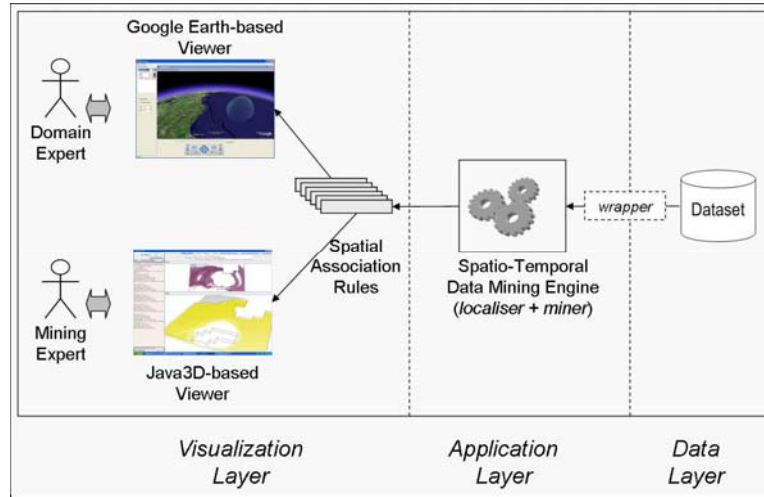


Figure 1: System Architecture

To interpret the output of the mining process, we envision feeding the results of the mining process to different visualization tools, possibly providing complementary interacting functionality. In the current implementation, we have developed two alternative visualization tools to support exploratory visual data interpretation. The first one (described in Section 4.1) embeds the Google Earth application while the second one (Section 4.1.2) is a Java3D-based application. Both applications are able to display the output of the mining process in a 3D virtual environment, allowing the user to freely change their viewing perspective. The main goal of these applications is to enhance the overall knowledge discovery process, allowing decision makers and knowledge engineers to better understand and discuss the logic behind the models. Since several different application domains can be considered, the application server must include domain-specific wrappers that transform raw data into the input format required by the mining engine.

We propose a new approach for spatio-temporal data mining, consisting of two main components: *localiser* and *miner*. The *localiser* deals with the data attributes and especially with spatial and temporal dimensions. The *miner* processes the data based on the spatio-temporal relationships provided by the *localiser*. By differentiating the two tasks, the computations are simple and quick as they are done locally. The output of the *miner* (association rules) is also fed to the *localiser* for further improvement. Unlike conventional approaches, in which the interactivity is inexistent or reduced to a minimum (very few simple operations), our approach is interactive as the two processes can run concurrently. Moreover, in this approach the processes of mining/analyzing the data and visualization are quite separate, as to

facilitate the testing and evaluation of different algorithms and techniques involved in each individual phase.

3. Spatio-temporal Data mining

Usually the pre-processing phase depends directly on the technique used to mine and analyze the data. In this study, we developed a technique to categorize data based on the sampling technique described in (Toivonen, 1996). The main difference between them resides in their goals. Our technique is designed to properly categorize the variables or attribute values, while the sampling technique developed in (Toivonen, 1996), its main goal is to reduce the database activity by analyzing a randomly chosen sample and then generalizing the result to the whole database. Our technique also takes into account the fact that many variables never cover all the range of values allowed. Thus, the model used here consists of mapping the spatial datasets onto a virtual partitioned space. This can be seen as a layer in which original data is aggregated into virtual points (partitions) representing the minimal spatial unit that can be occupied by a spatio-temporal entity. Each virtual point is identified by a set of attributes including coordinates, size, neighborhood, etc. For instance, traditional geographical databases have two or three dimensions, while in spatio-temporal datasets the number of dimensions can range from two (one spatial and one temporal dimension), to N (time, three spatial dimensions, n virtual dimensions). The points are disjoint; therefore, any shape used to implement a virtual point should satisfy disjunctive and complement properties. This model will allow us to hide all the problems of heterogeneity and unify the concept of items (virtual points) to the majority of spatio-temporal datasets.

In the proposed system we focus our attention on developing a technique based on association rules to discover relationships between spatial patterns. A spatial association rule is of the form “ $A \rightarrow B$ (s%, c%)”, where the pattern A is called *antecedent* and B *consequent*, and the percentages s and c are called the *support* and the *confidence* of the rule. The problem of discovering association rules consists of identifying all rules, within the dataset, satisfying minimum support s and confidence c. This usually requires a solution to the following two sub-problems: 1) find frequent (large) spatial patterns; 2) extract strong spatial association rules. In the first problem the rules should satisfy a minimum support (support > s) and in the second a spatial association is said to be strong if it satisfies a minimum confidence (confidence > c).

To mine spatial association rules we developed a technique based on Apriori algorithm (Orlando et al., 2001), which is based on the rule: “*any subset of a frequent itemset must be frequent*”. The association rule extraction is based on the key concept of spatial itemset. According to the model defined above for spatio-temporal datasets, each itemset is associated to a set of virtual points. We say that a virtual point *supports* an itemset if and only if the itemset is frequent in that point. That itemset is called spatial itemset. Note that virtual points supporting the rule (or itemset) can cover more than one time step, thus extending the model to include patterns with well-defined time interval. As the traditional Apriori algorithm has a very high computational complexity, it is not suitable for very large datasets. The idea is to reduce the size of the input data by presenting to the algorithm only data with higher spatio-temporal relationship; namely virtual points. Therefore, by exploiting the features of spatio-temporal datasets and by reducing the size of candidate generation performed by the localiser, the adapted Apriori algorithm is efficient. Basically, two itemsets can be united if they share all the items except one. A further control is needed to verify that the intersection of the sets of virtual points referred by the two itemsets is not empty. Indeed, such intersection will become the supporting area (set of virtual points) for the new itemset. This is to ensure that each new itemset has a supporting area, and an adequate number of virtual points. The output of the algorithm consists of frequent itemsets and strong association rules.

4. Visual Techniques for Advanced Spatial Analysis

In the following, we describe the two visualization applications we developed in order to provide our system with various exploratory visual capabilities, meant for the different actors involved in the interactive mining process. Indeed, these tools allow to answer different but complementary requirements posed by domain and mining experts. While the Google Earth-based tool focuses on highlighting the spatial relationships between the dataset and the real-world geographical entities involved in the phenomenon, the Java3D-based tool mainly concentrates on the exploratory analysis of the data, by analyzing the internal structure of the dataset, their inherent internal relationships, and the patterns among data inferred by the mining algorithm. These two tasks are quite different, in terms of both the displayed information and the functionality required to explore and interact with the data. Thus, the combination of these two tools provides the user with very appropriate and effective means of studying the problem, while avoiding visual/cognitive overload (due to unnecessary rendered information, cluttering the

display) as well as limitations in exploratory analysis. Moreover, the system allows for validating both the spatio-temporal mining process and the discovered patterns.

4.1. The Google Earth-based tool

The first tool has been meant for domain experts, i.e., users that study the specific phenomenon but are not (necessarily) experts in data mining. To this aim, the tool exploits the 3D capabilities provided by Google Earth (shortly GE), where the user can freely move his/her perspective view in a 3D environment, to combine dataset themes and variables with real world infrastructures and geographic features. GE is a virtual globe combining satellite raster imagery with vector maps and layers, in a single and integrated tool, to allow users to interactively fly in 3D from outer space to street level views. A very wide set of geographical features (streets, highways, borders, rivers, etc.) and points of interest can be overlaid onto the map. The application uses data from NASA databases to render 3D terrain models, thus providing also Digital Elevation Model features. A key characteristic of this tool is the fact that the geographical data are not stored on client computers, as they are streamed, upon request, from Google's server infrastructure. This also guarantees that data are always up-to-date.

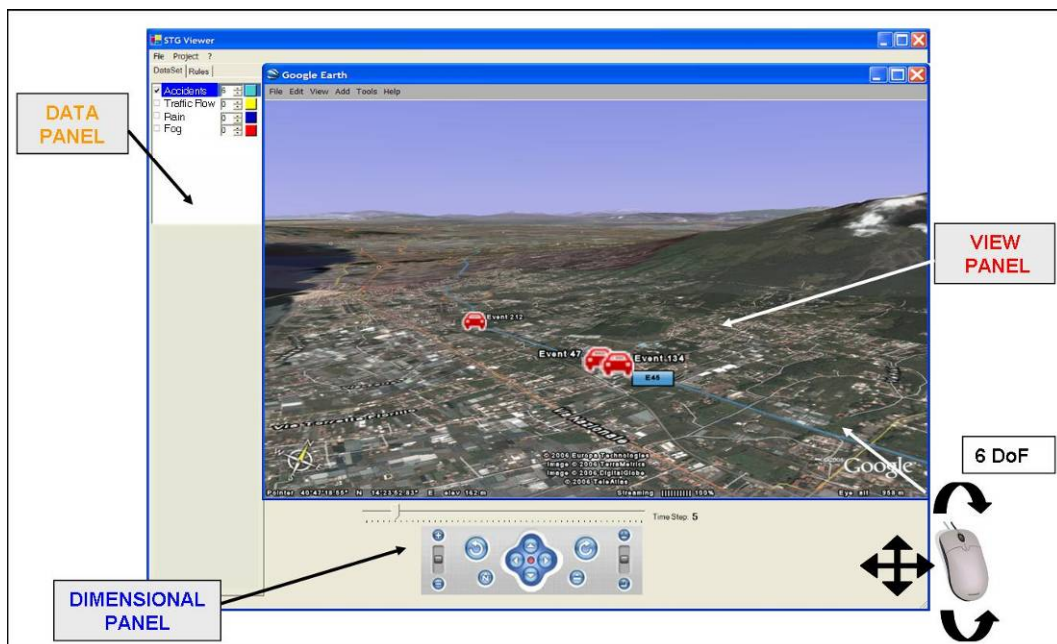


Figure 2: Schema of the User Interface for the GE-based visualization tool

The system we have developed embeds GE, which is used basically to render in 3D the information, while our application is responsible to allow the decision makers to select and/or filter the data to show, to customize the way they are represented (e.g.: icons, clouds of points, pie charts, etc.), and to move in the

temporal dimension. A screenshot of this application is shown in Figure 2, representing the considered Highway, and some events. By analyzing this figure, it is possible to notice that the resulting User Interface is composed of three main areas. The leftmost one (*Data Panel*) allows the decision maker to query the dataset, in order to define the specific (set of) themes/rules to render, and to customize the way the information is depicted. Through a Tabbed control, the user can choose if dealing with the attributes of the whole dataset, or with the association rules inferred by the mining engine. The bottom area (*Dimensional Panel*) allows the user to move in four dimensions, namely the 3D permitted by GE (by exploiting six degree of freedom), and the time dimension, through a sliding bar. Finally, the central, bigger area (*View Panel*) contains the GE application, used to show in 3D, at an arbitrary zoom level, the data.

Each event/variable is rendered as a specific icon or a cloud of points with a different color. The application works accordingly to the principle that all the data that do not match these query parameters, set by the user, are removed from the visualization canvas. This filtering is immediately applied, thus providing direct manipulation features. Moreover, each rendered point can be made an hyper link, in order to answer query of the form “When + Where \rightarrow What”, which is a typical analysis task in exploratory spatio-temporal data-mining (Peuquet, 1994). This application turns out to be very flexible, being able to deal with a large variety of spatio-temporal phenomena, ranging from worldwide (e.g.: weather, pollution, epidemic diffusions, etc.) to local ones (e.g.: local health, traffic, etc.).

From a technical point of view, we exploited the ad-hoc language provided by GE, named *Keyhole Markup Language*, or KML (KML Specifications, 2006), which is an XML grammar and file format suited to model one or more spatial features to be displayed in GE. We designed and implemented some procedures to generate on-the-fly the KML files, basing on user input specified in the Data Panel. Each generated KML file contains the coordinates of each considered point of the data (or item) set. Moreover, we used the set of API provided by GE to grant a full control on the User Interface and the active point of view in GE. Finally, it is worth noting that since KML files can be easily shared over the IP protocol, the system can be used for collaborative visualization and participatory decision making processes.

4.2. The Java3D-based tool

The Java3D custom application we have developed is aimed at providing a 3D rendering of and interaction with the association rules produced by the mining algorithm. Basically, it is a mining expert-

oriented tool, since it offers many features which are specific for the exploratory data mining domain. For instance, it provides some widgets on the interface to directly choose different values, parameters or Association Rules to display.

The visualization tool we have developed is apt to present in 3D the Association Rules identified by the mining engine. It exploits the “Arranging view” visualization technique (Andrienko et al., 2003), where two different views are presented in separated windows, and the user can arbitrarily arrange them to facilitate the comparison of data. Consequently, data are shown in two canvases, which can be rotated, zoomed and moved to easily examine shape, density and inner pits of the cloud of points. The User Interface, shown in Figure 3, is composed of six main panels, suited to select the rule to display among the ones inferred by the mining engine, to set the required level of confidence, and to choose which timestep the data has to be fetched from. Moreover, two Java3D canvases are aimed at rendering the selected (active) antecedent and consequent of the current rule. These canvases can be freely resized: for example, one can be closed, in order to allow the maximum flexibility and customization of the visualization space. For further details on the user interface and on the provided features, see (Compieta et al., 2006).

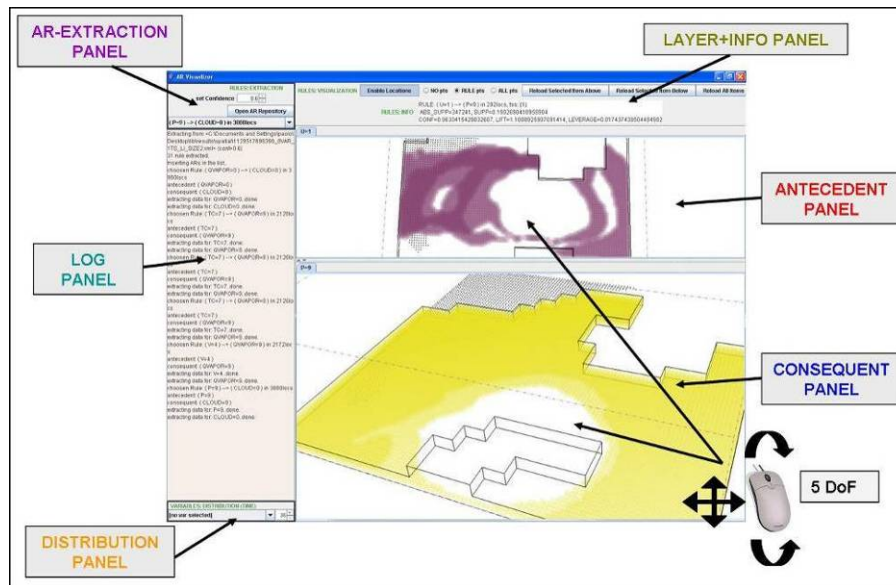


Figure 3: Schema of the User Interface for the Java3D-based visualization tool

The main strength of this application is the innovative and designed-on-purpose functionality of drawing, upon user request, the “shape” of a rule, intended as that particular region of the space where the rule holds – that is, the set of locations in which the rule (hence all item involved in it) are well supported.

While simply removing confusion and overload of visual information from the screen, it also help to highlight the structure of any pattern embedded in the data and to focus the user's attention only on the subset of the dataset involved in the rule being studied. This allows a more efficient and light visualization process, even when displaying millions of points. It is clear that not only the content of a rule is important, but also its shape: from a domain-expert's point of view, it might tell lots of information about the behavior of the phenomenon being studied – thus being able to narrow the visualization phase only to that shape is extremely valuable in interpreting all results. To the best of our knowledge, the standard visualization tools for geo-spatial data do not provide this functionality.

5. Ongoing work

We have developed a system for mining spatio-temporal datasets. The system has been initially develop to handle punctual data and specifically tailored towards the case study selected. However the architectural design of our system is very flexible and allows the implementation to be adapted to work with other datasets: indeed it is independent of both the specific data and the particular mining algorithm utilized. As we have obtained very interesting results (in terms of knowledge discovery and exploratory visualization) with our case study, we now intend to use our system for road safety applications. To this aim we have obtained a massive dataset from the Italian highway agencies, containing information of events (especially fatal accidents) occurred along a given stretch of highway over the last three years. This data differs from the data previously utilized as it contains not only punctual data but also linear data. Therefore our mining techniques must be adapted to take this characteristic into account. Moreover, the dataset is very heterogeneous, including information on the morphology of the terrain, all significant events occurred on the highway, traffic conditions, and meteorological phenomena. It contains a very detailed map of the highway, subdivided into segments. For each segment, number of lanes, speed limit, kind of tarmac, presence/absence of hard shoulder (and its width), etc., are specified, together with surrounding objects (e.g., gas stations, road signs, lay-by, etc.). The road traffic events, the status information and some of the meteorological events are modeled using the DATEX standard, which defines the logical data structure and the terminology used. This standard presents also a common Data Dictionary, a common set of messages based on the ISO standard EDIFACT (Electronic Data Interchange for Administration, Commerce and Transport) and a common Geographical messaging system. In particular, for each event, the database contains an ID, a description based on some metadata, the

geographical coordinates, and a set of attributes, specific for the occurred event. It also contains information about the traffic flow including number of vehicles that entered and exited the highway. Moreover, for each vehicle equipped with “Telepass” (a wireless system for automatic toll), the timestamp of the entry and exit stations is stored. Finally, data collected from sensors placed in the tarmac, able to log the number of passing vehicles, the type (car, truck, etc.) and speed of each vehicle, as well as the distance between consequent vehicles are recorded. Some of the meteorological data (e.g., wind, fog, etc.) are also modeled with the DATEX standard, but the dataset contains also information coming from the rain gauges placed around the track.

We are currently pre-processing the Gigabytes of available data. We are planning to embed new algorithms in our data mining engine, to compare them and to identify the most suited for the specific characteristics of this dataset. Similarly, from a visualization point of view, we are investigating the user of metaphors to represent the different and heterogeneous information of the dataset. Furthermore, we are improving our GE-based application to provide more customizations options for data representation, while we intend to integrate cartographical layers within the 3D canvases of our Java3D application.

References

- Andrienko N., Andrienko G., and Gatalsky P., Exploratory Spatio-Temporal Visualization: an Analytical Review. *Journal of Visual Languages and Computing*, special issue on Visual Data Mining. December 2003, v.14 (6), pp. 503-541.
- Andrienko N., Andrienko G., “Exploratory Analysis of Spatial and Temporal Data – A Systematic Approach”, Springer, 2005.
- Bédard, Y. Spatial OLAP. 2ème Forum annuel sur la R-D, Géomatique VI: Un monde accessible, 1997, Montréal, CA.
- Bédard, Y., Merrett, T., and Han, J. Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery. *Geographic Data Mining and Knowledge Discovery*. Taylor & Francis, London, 2001, 53-73.
- Compieta P., S. Di Martino, M. Bertolotto, F. Ferrucci, T. Kechadi. Exploratory spatio-temporal data mining and visualization. To appear in *Journal of Visual Languages and Computing*, Special Issue on Visual Languages and Techniques for Human-GIS Interaction, 2006.
- Costabile M.F., Malerba D. (Editors), Special Issue on Visual Data Mining, *Journal of Visual Languages and Computing*, Vol. 14, December 2003, 499-501

DATEX Standard Specifications, 2005. Available at <http://datex.eu.org/> , last visited July, 03, 2006.

Fayyad U.M., Grinstein G.G., Introduction, in: Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, Los Altos, CA, 2001, pp. 1–17.

Google Earth web site, 2005, available at: <http://earth.google.com/>, last visited July, 03, 2006.

Java3D web site, 2001, available at: <http://java.sun.com/products/java-media/3D/> , last visited July, 03, 2006.

Johnston W.L., Model visualization, in: Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, Los Altos, CA, 2001, pp. 223–227.

KML Specifications, available at http://earth.google.com/kml/kml_intro.html , last visited July, 03, 2006.

Kopanakis I., Theodoulidis B. Visual data mining modeling techniques for the visualization of mining outcomes. Journal of Visual Languages and Computing. 14(6): 543-589 (2003)

National Hurricane Center, 2003. “Tropical Cyclone Report: Hurricane Isabel”, <http://www.tpc.ncep.noaa.gov/2003isabel.shtml>, last visited July, 03, 2006.

Orlando S., Palmerini P., and Perego R. Enhancing the Apriori Algorithm for Frequent Set Counting. In Proc. of 3rd Int. Conf. on Data Warehousing and Knowledge Discovery (DaWaK 01) , volume 2114 of LNCS, pages 71-82. Springer, 2001.

Peuquet D.J., It’s about time: a conceptual framework for the representation of temporal dynamics in geographic information systems, Annals of the Association of American Geographers 84 (3) (1994) 441–461.

Shneiderman B., Inventing discovery tools: combining information visualization with data mining, Information Visualization Volume 1, Issue 1 (March 2002), Pages: 5 - 12 ISSN:1473-8716

Toivonen H., “Sampling large databases for Association Rules”, Proceedings of the International Very Large Database Conference, pp.134-145, 1996.

Zhang T., R. Ramakrishnan, and M. Livny. BIRCH: An efficient data clustering method for very large databases. In SIGMOD, 1996.