Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Spatio –Temporal Data Mining

[1]Radha K., [2]Sri. Balaji G
[1]M. Sc, M. Tech, B. Ed, PHd (Registred) Assoc. Professor,
KKC Institute of Technology & Engineering, Puttur, Karnataka, India
[2]M. Sc, M. C. A, M. Sc(Psy), B. Ed, M. Tech (CSE) Assistant Professor,
KKC Institute of Technology & Engineering, Puttur, Karnataka, India

Abstract: ***Spatio temporal data sets are often very large and difficult to analyze and display. Since they are fundamental for decision support in many application contexts, recently a lot of interest has arisen toward data-mining techniques to filter out relevant subsets of very large data repositories as well as visualization tools to effectively display the results. In this paper we propose a data mining system to deal with very large spatio -temporal data sets. Within this system ,new techniques have been developed to efficiently support the data-mining process ,address the spatial and temporal dimensions of the data set, and visualize ,interpret results .In particular two complementary 3D visualization environments have been implemented .One exploits "Google Earth" to display the mining outcomes combined with map and other geographical layers, while the other is a Java 3D-based tool for providing advanced interactions with the data set in a non-geo-referenced space, such as displaying association rules and variable distributions.***

Keywords: Data mining, Spatio -temporal data, Exploratory visualization, 3D visualization.

I. INTRODUCTION

During the last decade , our ability to collect and store data has far outpaced our ability to process, analyze and exploit it. Many organizations have begun to routinely capture huge volumes of historical data describing their operations, products and customers.At the same time scientists and engineers in many fields have been capturing increasingly complex experimental data sets,such as terabytes of data received daily from space-borne instruments, high spatial, temporal and spectral-resolution remote sensing systems, and other environmental monitoring device .Some researchers estimate that about 80% of the data stored in corporate databases integrate spatial information , leading to huge amounts of geo-referenced information that need to be analyzed and processed. These data sets 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 mining and analyzing spatio-temporal data sets has recently emerged as a research priority.

Data mining techniques have been proven to be of significant value for spatio-temporal applications. It is a user-centric, interactive process where data mining experts and domain experts work closely together to gain insight on a given problem. Inparticular spatio-temporal data mining is an emerging research area, encompassing a set of exploratory , computational and interactive approaches for analyzing very large spatial and spatio-temporal data sets. 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 presenting the final results.

Visualization techniques are widely recognized to be powerful in this domain, since they take advantage of human abilities to perceive visual patterns and to interpret them. To address these issues we have developed a system for exploratory spatio-temporal data mining.The aim of this system is on one hand to enable data-ming 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, thus leading to greater effectiveness and significance of the results.To achieve these goals the system includes a data mining engine that can integrate different data mining algorithms and two complementary 3D visualization tools.

II. SYSTEM ARCHITECTURE

This section describes the architecture of the system. we first discuss the main concepts of the data-mining process, and then introduce the main components of the system ,namely the mining engine and the visualization tools.

2.1 The spatio-temporal data-mining process

The data mining process usually consists of three phases , or steps
1) Pre-processing or data preparation. 2) Modeling and validation 3) Post processing or 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.

The third step of using this model evaluated and validated in the second phase, to effectively study the application behavior. The mining process for spatial data is more complex than for relational data in terms of both the mining efficiently and the complexity of possible patterns that can be extracted from spatial data sets. Therefore, new techniques are required to efficiently and effectively mine spatial data sets. Especially in spatial data mining the third phase is so important that some researchers incorporatedmost of its processes into phase two such as automatic and interactive visuzlization of data,and called it interactive data mining(IDM).

2.2 Exploratory spatio-temporal data-mining system

The proposed system for mining large spatio-temporal data sets describes the behavior of some natural phenomena, which have been monitored and recorded at several time instants. Our system relies on a standard three-tier architecture ,including a data store at the back end,an application server,and two visualization components at the front end.

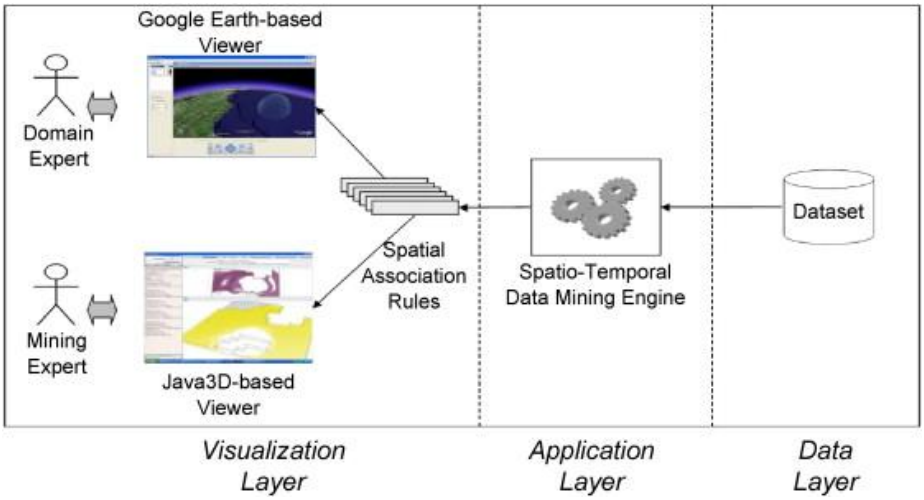

Fig. 1. System architecture.

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. These wrappers implement the data set type models described in

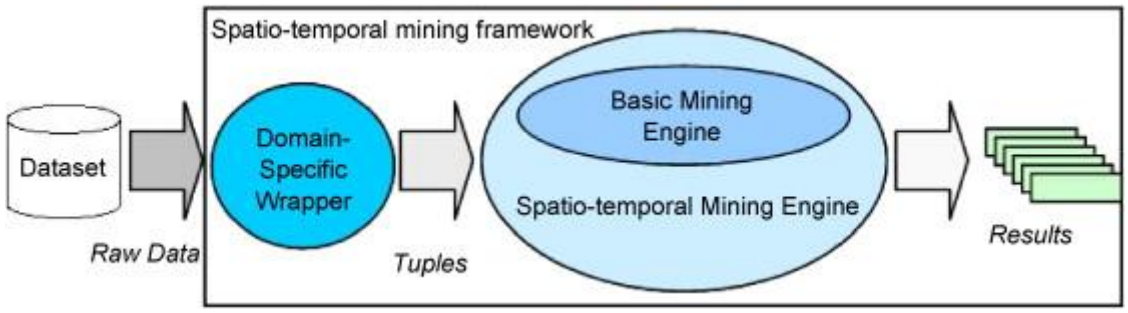

Fig. 2. Architecture of the spatio-temporal data-mining engine.

III. THE DATA-MINING PROCESS

In this section we describe our data-mining approach that deals with spatio-temporal data sets.We start with a review of the current state-of the art in this field, and then we present our solution.

3.1 Related work on spatial data mining

Numerous research projects on spatial data mining have been conducted in the last two decades. Some attentions have been dedicated to the application of existing as well as the development of new mechanisms to extract relevant information from large data repositories. However, due to the huge volume and diverse nature of this kind of data, traditional techniques such as statistical methods have high computational burdens and seem often inadequate to elicit complex spatial and temporal relationships among data. Association rules have also been used successfully on special data sets. The main idea is to design spatial association rules that not only can find local correlations between patterns ,but also global ones .Spatial association rules constitute an improvement to generalization-based spatial data-mining methods, as they cannot discover rules reflecting spatial pattern structures.

Most efforts have been spent in trying to adapt, modify or improve conventional techniques ,relying on a solid knowledge discovery in database(KDD) experience to design new suitable mining models. Some good attempts have been made in proposing a better-more meaningful-data format to highlight spatio-temporal relationships prior to elaboration; this can be achieved either by preprocessing the database or by imposing a level of meta-data to properly access information within the whole data set.

3.2 The proposed approach

In this paper we propose a new approach for spatio-temporal data mining,whose conceptual schema is depicted in Fig.3. The approach consists of two main components;localizer and miner.the localizer deals with the data attributes and especially with spatial and temporal dimensions. The miner process the data based on the spatio-temporal relationaships provided by the localizer.

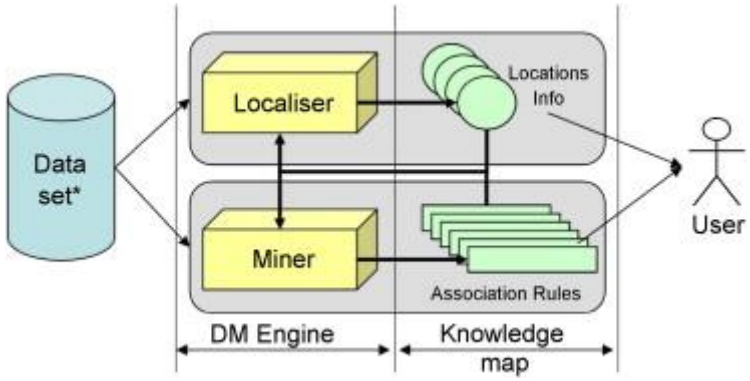

Fig.3.A schematic view of the proposed approach for spatial data mining.In the following we will focus on the techniques used in each phase.

3.3 Spatio-temporal data set model

It was already mentioned above that spatial data sets more complex than conventional data.
This complexity is not only in processing and interpreting the data but is also present at the data- mining process inputs . spatial data is usually characterized by two different types of attributes : spatial and non-spatial attributes. The former identifies the spatial locations of spatial items. These include 3D space coordinates, item shape, temporal, geometry, etc .The latter is usually the same as in conventional data sets such as item name , item key ,type ,rate ,size ,etc .The main difference between these two types of attributes is that the relationships between spatial patterns/items are often implicit ,while they are usually explicit in non-spatial objects. Some methods of how to optimize the categories have been developed and ,mainly based on heuristic methods and clustering analysis.

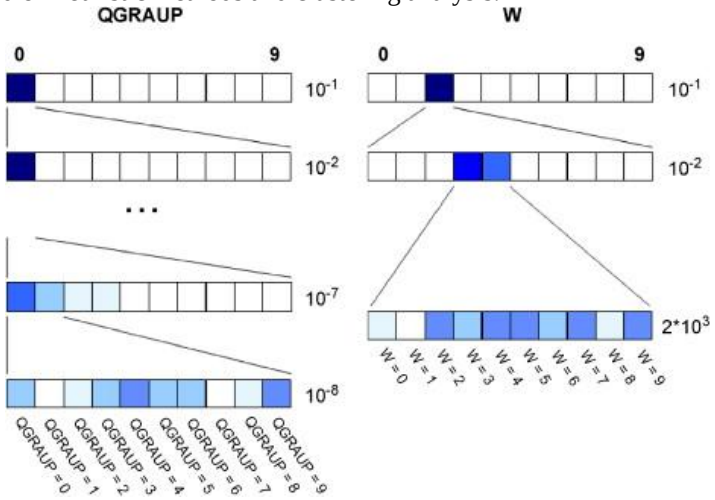

Fig.4.Example of categorization for the two variables W and QGRAUP. Dark color=many points are inthat range of values;light color=few points.

The model used here consists of mapping the spatial data sets onto a virtual partitioned space.this can be seen as a layer in which original data are aggregated into virtual points representing the minimal spatial unit that can be occupied by a spatio-temporal entity.Each virtual point is defined by a set of attributes including coordinates,size,neighborhood,etc.For instance,traditional geographical databases have two or three dimensions,while in spatio-temporal data sets the number of dimensions can range from two to N.

3.4 Spatial association rules

In the proposed system we focus our attention in developing a technique based on association rules to discover relationships between spatial patterns. A spatial association rule is of the form "A 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 data set, 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).

IV. VISUAL TECHNIQUES FOR ADVANCED SPATIAL ANALYSIS

Visual data mining refers to methods, approaches and tools for the exploration of large data sets by allowing users to directly interact with visual representations of data and dynamically modify parameters to see how they affect the visualized data.

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.

Google Earth (shortly GE) is a virtual globe, currently freely available for personal use on PC running on Windows and Mac OS, while the Linux version is expected shortly. For commercial and professional use, many purchasing options are available, ranging from basic licenses to enterprise services. The original project was developed by *Keyhole*, which was bought by Google in 2004.

Google Earth combines satellite raster imagery, with vector maps and layers, in a single and integrated tool, which allows users to interactively fly in 3D from outer space to street level views. Most places of the world are available at (at least) 1 km of resolution, while many large cities are available at high enough resolution to see individual buildings, houses, and even cars. A very wide set of geographical features (streets, borders, rivers airports, etc.), as well as commercial points of interest (restaurants, bars, lodging, shopping malls, fuel stations, etc.), 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.

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, economics, etc.). The tool we developed embeds GE and presents the same ease of use, resulting very suitable for domain-expert users.

The resulting user interface, shown in Fig. 6, is composed of three main panels:

View panel: This is the GE application, used to show in 3D, at an arbitrary zoom level, the data. It allows for six degree of freedom (DoF), achieved through combination of mouse clicks and movements, or through a lower panel.

Data panel: This is a vertical panel, located to the left of the window, which allows to query and filter data, to get the specific information that the user wants to view. Through a Tabbed control, the user can choose if dealing with the attributes of the whole data set (the first tab), or with the association rules inferred by the mining engine (the second tab).

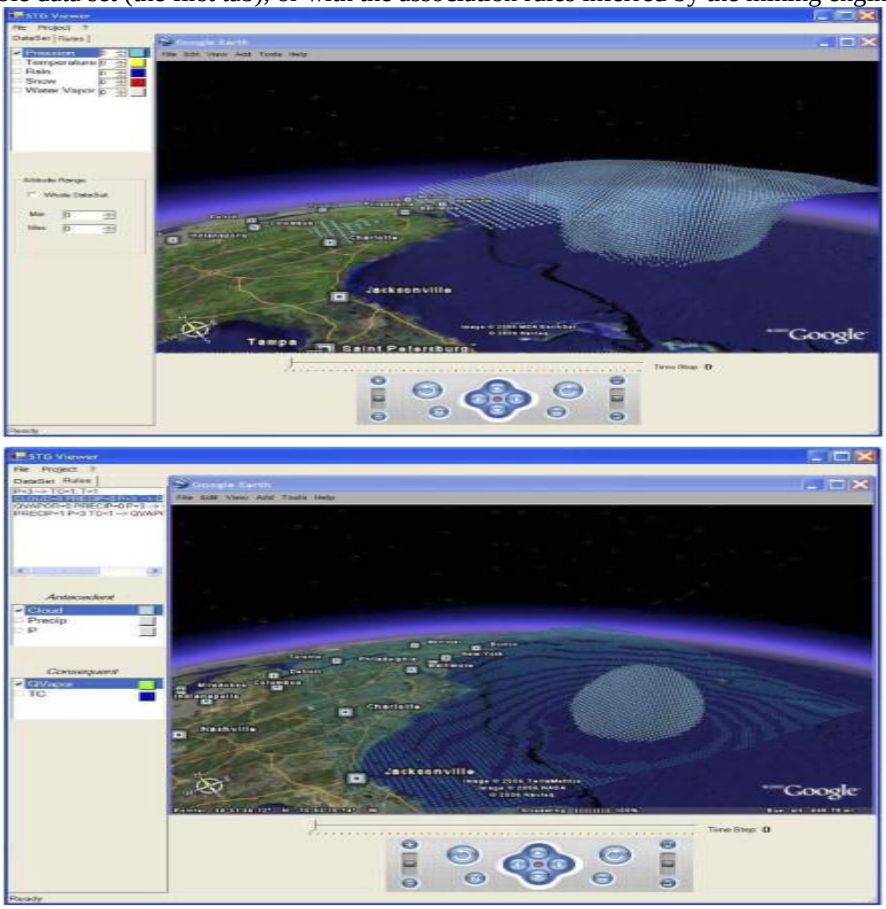

Fig. 7. Visualization of data set themes (Top) and rules (Bottom). Notice the different widgets provided in the left panel.

Dimensional panel: This panel allows the user to move in four dimensions, namely the 3D permittd by GE (by exploiting six DoF), and the time dimension, through a sliding bar. To this aim, the horizontal panel, located at the bottom of the window, realizes a unique control panel/set of commands to follow the data painted on screen.

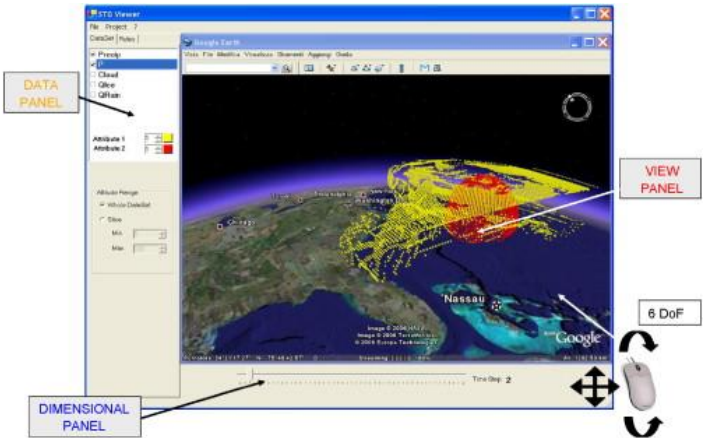

Fig. 6. Schema of the user interface for the GE-based visualization tool.

4.1.1. Technical Aspects

The development of an environment exploiting Google earth technologies to render mining outcomes posed two main challenges:

1) How to arrange the information of the data set and/or the rules in a way that it could be displayed by Google Earth.
2) how to improve the Google Earth user interface to provide the tools to carry out the exploratory spatio-temporal data mining and visualization.

The first issue we exploited the ad hoc language provided by GE, named keyhole markup language (or KML),which is an XML grammar and file format suited to model one or more spatial features to be displayed in GE. In relation to the second issue, there are two main ways to programmatically interact with GE. This approach is straightforward, but does not provide an effective management of the user interactions. The alternative solution is to use the set of API provided by GE. Indeed, once such an application is installed, a new COM component is available in the Windows system, namely the ***KEYHOLEib*** . Once this is imported in a programming environment as a reference library ,a new namespace is available ,which exports two main classes ,i.e. the ***KHInterface*** Class and the ***KHViewInfoClass***.

4.2. The Java 3D-Based Tool

The java 3D custom application we have developed is aimed at providing a 3D rendering of and interaction with the association rules produced by the mining algorithm.

The user interface, composed of six main panels:

AR-Extraction Panel: This panel extracts the rules(in XML format) from the output of the mining algorithm.it contains a field to set required level of confidence, a button to choose the file, and a combo box to choose,among the rules contained in the selected file,the one to display.

Layer Info Panel: This panel allows to set rendering styles and shows detailed information about the rules. It allows the user to chose whether to visualize the 'land'(ground) or 'locations'(association rule's bounding cubes-the supporting area) layer.

Log Panel: This panel is aimed at providing some textual output to the user-general information on the execution state of the drawing and data-fetching threads.

Antecedent Panel: It is a JAVA 3D canvas, aimed at rendering the selected(active) antecedent of the current rule.The active antecedent can be chosen through a set of tabbed controls,each corresponding to a tabbed panel,placed on the top of the canvas.This canvas has five DOF,related to mouse movements and button clicks.

Consequent Panel: As the previous panel, an itemset is shown through several tabbed inner panels as well as there are five DOF,related to mouse movements and button clicks.

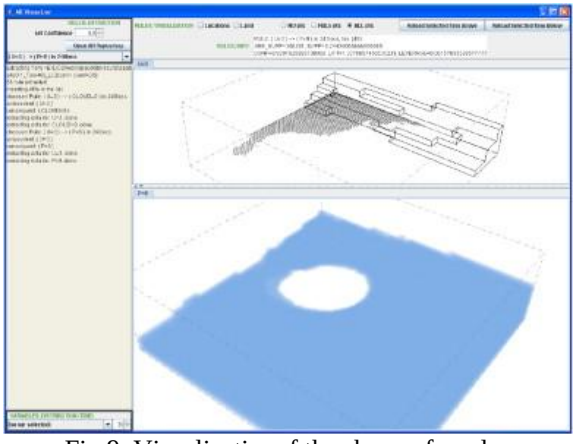

Fig.9. Visualizationof the shape of a rule.

Distributed Panel: This panel allows to render of the data set(e.g."Pressure"),independently of the rules involving it. Each values 0,1,……..9 distribution,at a given timestep is displayed separately.

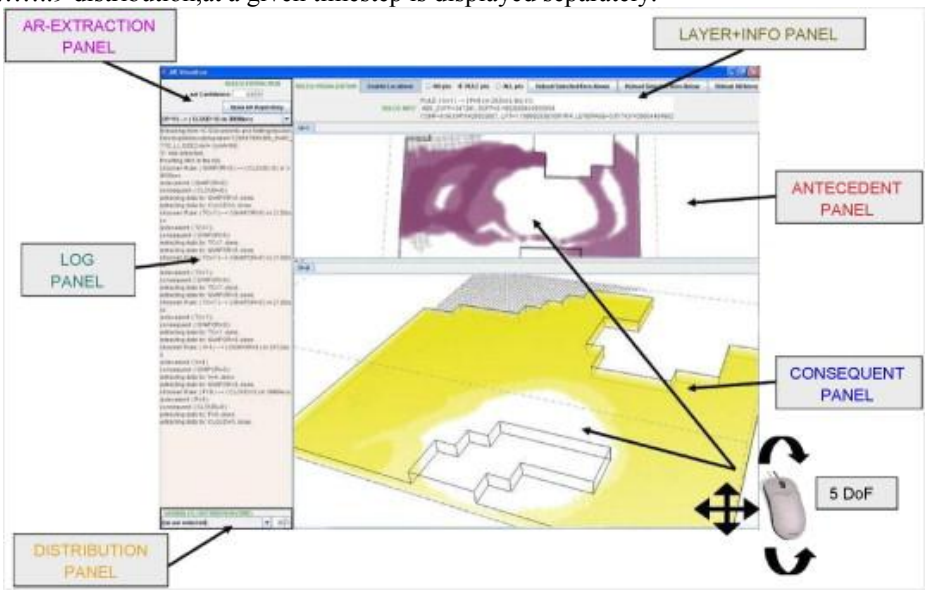

Fig.8.Schema of the user interface for the Java 3D based visualization tool.

4.2.1. Technical Aspects:

The development of this application required us to design and implement a system exploiting many different technologies, including 3D computer graphics, RDBMS connections, XML, etc.For each of these technologies we have surveyed existing high-level, versatile solutions. For instance,many different programming libraries for delivering 3D applications are currently available, including OpenGL,Direct3D and Java3D.Similerly,many alternatives exist to connect to RDBMS(e.g. ADO ,JDBC ,etc). We choose to develop our application within the Java programming environment,whose key advantage is the large availability of coherent ***Application Programing Interface***(API),delivered by ***Sun Microsystems.***

V. A. CASE STUDY: HURRICANE ISABEL:

We have tested our system on a large spatio-temporal data set. This sections detail the data format and the experimental results obtained.

5.1. The Data Set

Hurricane Isabel was the only Category 5 hurricane of the 2003 Atlantic hurricane season(see Fig:10).It made Landfall on September 18,2003 in North Carolina. Official reports state that an official damage estimate of 3.37 billion of US Dollars.

Fig.10.Hurricane Isabel shot,from satellite

All the key data about this phenomenon were logged for two days by the National Center for Atmospheric Research in the United States. The corresponding data set was produced by the Weather Research and Forecast(WRF) model,courtesy of NCAR,and the U.S.National Science Foundation(NSF).

Variable	Descripton	Min/Max	Units
QCLOUD	Cloud Water	0.00000/0.00332	kg/kg
QGRAUP	Graupel	0.00000/0.01638	kg/kg
QICE	Cloud Ice	0.00000/0.00099	kg/kg
QRAIN	Rain	0.00000/0.01132	kg/kg
QSNOW	Snow	0.00000/0.00135	kg/kg
QVAPOR	Water Vapor	0.00000/0.02368	kg/kg
CLOUD	Total cloud (QICE+QCLOUD)	0.00000/0.00332	kg/kg
PRECIP	Total Precipitation (QGRAUP+QRAIN+QSNOW)	0.00000/0.01672	kg/kg
P	Pressure; weight of the atmosphere above a grid point	-5471.85791/3225.42578	Pascals
TC	Temperature	-83.00402/31.51576	Degrees Celsius
U	X wind component; west-east wind component in model coordinate, postive means winds blow from west to east	-79.47297/85.17703	
V	Y wind component; south-north wind component in model coordinate, postive means winds blow from south to north	-76.03391/82.95293	
W	Z wind component; vertical wind component in model coordinate, positive means upward motion	-9.06026/28.61434	

Fig.11.Hurricane Isabel Data set's Layers.

All variables are real-valued and were observed along 48 timesteps(once every hour for two days) in a space having $500*500*100=25*10^6$ total points. Each variable in each timestep is stored in a different file, resulting in 624 files of about 100 MB each .This fine fragmentation allows great flexibility in choosing different subset of data for each mining task.

Therefore the Hurricane Isabel Data set is proper instance of a massive geographical spatio-temporal data set, and is widely adopted for data visualization studies, such as the ones made for the IEEE 2004 Visualization Contest.

5.2. Visual results for exploratory analysis and decision support

In this section we show results obtained by applying the adapted Apriori algorithm to the Hurricane Isabel Data sets and by viewing them with the JAVA 3D tool described in some other places. We analyzed many results and report here only some of the most meaningful. In our analysis we tried to discover specific patterns/characteristics that were either well known about hurricane data. Having thousands of locations in the entire space permits millions of possible different shapes for a rule.While analyzing results the decisive attribute in the early screening has really been the area they covered, ,rather than their content. However, sometimes it turned out to be misleading: Fig.12 illustrates an example of apparently trivial shape that instead is interesting.

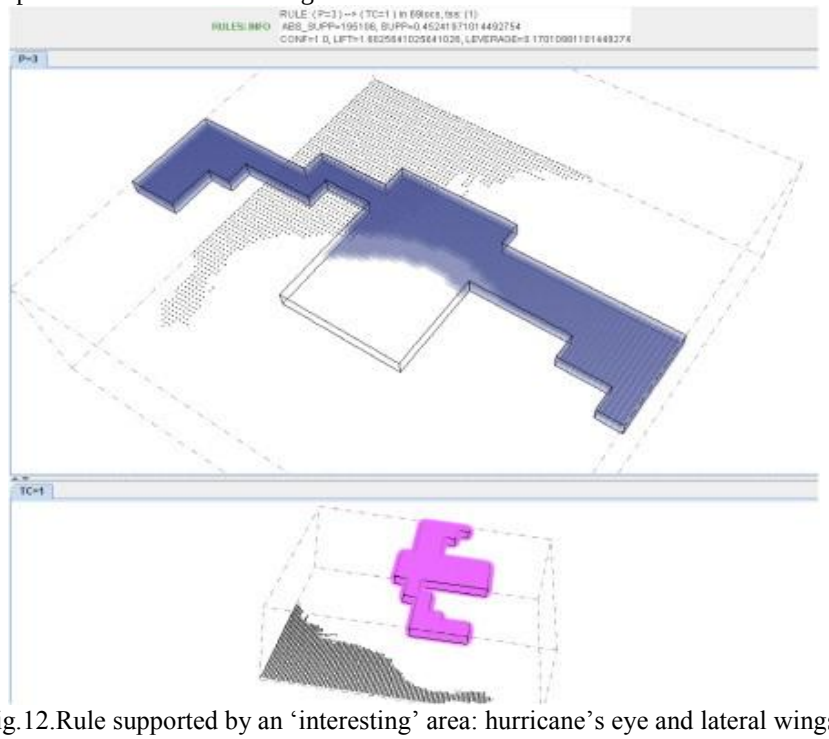

Fig.12.Rule supported by an 'interesting' area: hurricane's eye and lateral wings.

Indeed,fig.12 shows a rule covering a relatively small area corresponding to the eye of the hurricane, with two little backward-facing wings. It is well known that pressure assumes always its lowest value in the eye of the hurricane : The eye of the hurricane has much importance when predictions have to be made regarding its strength and speed. Therefore, rules featuring unusual, oriented or hurricane-shaped areas-even if including only a few locations-many be essential to discover new behaviors.

All variables represent natural events, each with a different, and often not well defined, distribution.Fig;13 illustrates an example :the distribution of the wind variable appears almost everywhere, although with very low support.

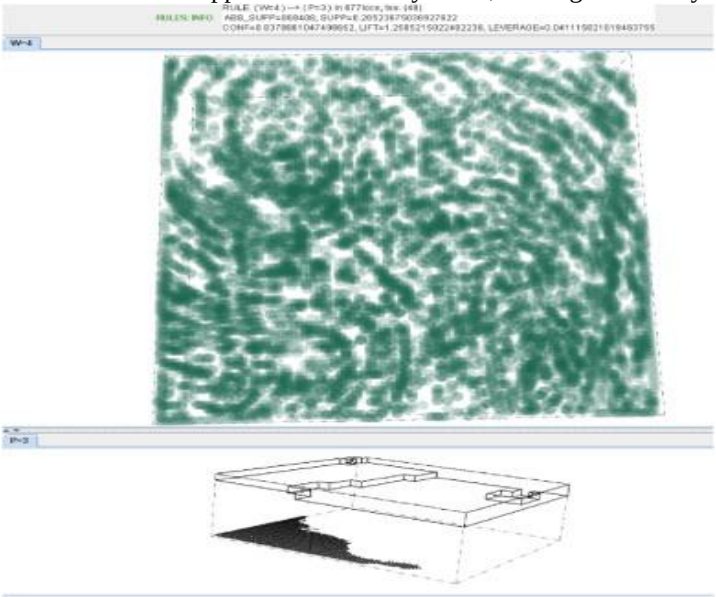

Fig.13.(w=4 p=3)(w=4 p=3) in ts=48.w is the wind's East-West component.Only the area supporting this rule is shown below.

When using photographs to analyze a hurricane ,all wind components are ***invisible;*** they Can only be guessed by the user and thus they are not exploited in a human study; on the other hand ,a normal mining process can detect such a presence but is ***incapable of localizing it or taking into account its 'density' with respect to space.*** In this situation, the method developed during this work becomes very effective: it can detect only those areas reporting a sufficient number of points having large values for that phenomenon. Actually, this process of selection applies a light form of ***compression*** to the data describing the event, that is much faster than ***punctual data mining*** he main objective in studying hurricanes is that of predicting what they are going to do in the immediate future.

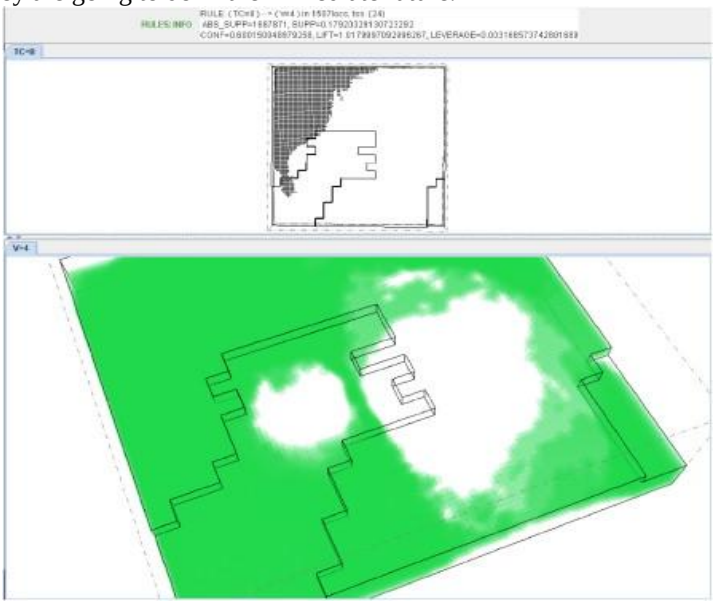

Fig:14.Rule(TC=8→V=4)(TC=8→V=4) in ts=24

By analyzing the results, we found that Cloud Water and Water Vapour were the parameters most affected by the seashore. This is clearly visible in the following two figures .In particular in Fig.15 the data points displayed in yellow represent the locations of the data set where the Cloud Water has a low level. Points in blue correspond to the eye of the hurricane. From this figure it is possible to notice how the altitude with a low level of Cloud Water suddenly drops, as soon as the phenomenon impacts the dry land.

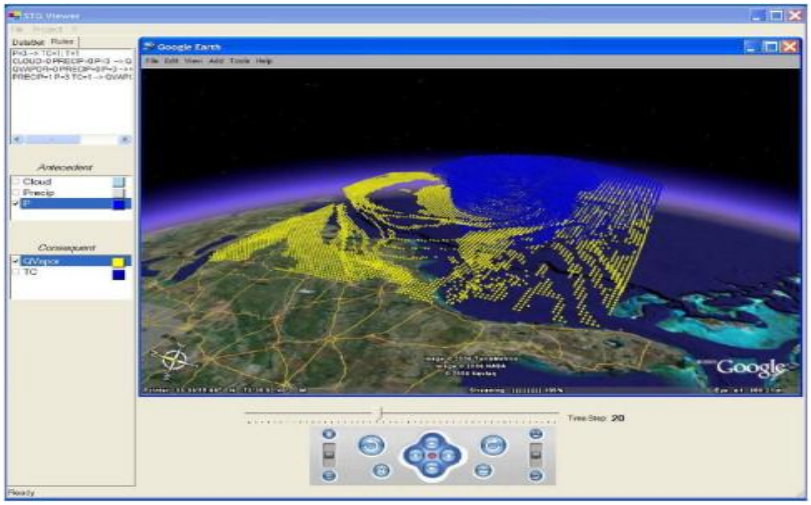

Fig:15. Effects of the dry land on the Cloud Water

A similar phenomenon is visualized in Fig.16.Yellow points represent the location with a high percentage of water vapour .It is evident that there is a cluster of points over the sea, but when the hurricane meets the land, these points almost disappear.

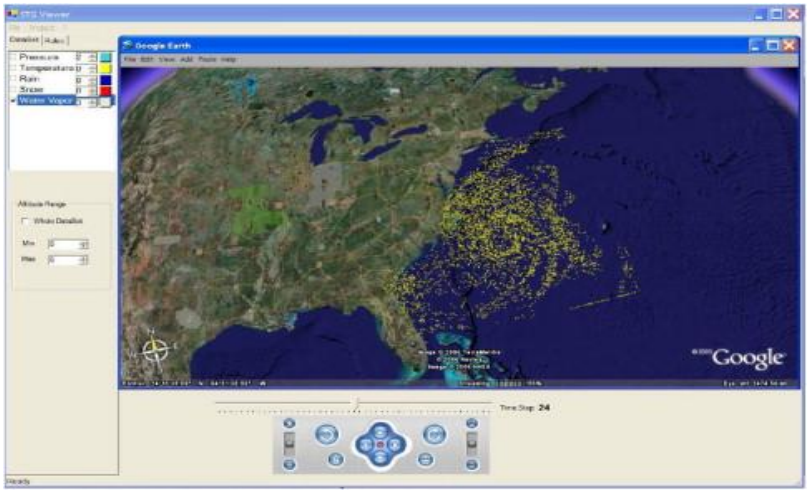

Fig:16 Effects of the dry land on the Water Vapour.

VI. CONCLUSION

In this paper we have described the system for exploratory spatio-temporal data mining we have developed. This system includes a mining engine based on an adapted version of the well-known A-priori algorithms. Since results of a mining algorithm require interpretation ,we have focused on visual techniques. To this aim we have developed two independent visualization tools for viewing and interacting with the results of the mining process, meant, respectively, for the domain experts and data mining experts.The Google Earth application allows to relate the phenomenon being studied to the specific geographic area and associated features.The second visualization tool presents more sophisticated interactively.Our system has been tested on a large real-world data set and has produced interesting results.However we plan to perform more extensive testing with domain experts.The system offers much scope for enhancements and further developments.We also intend to integrate the two visualization tools allowing to switch in a continuous fashion between them maintaining the same perspective.

REFERENCES

[1] R.Agarawal ,T.Lmielinski,A.Swami,Mining association rules between sets of items in large databases,in:ACM SIGMOD Conference,1983.

[2] U.M.Fayyed,G.G.Grinstein,Introduction in Information Visualization in Data Mining and Knowledge Discovery ,Morgan Kaufmann,Los Altos,CA,2001,PP.1-17.

[3] Y.Bedard,T.Merrett,J.Han,fundaments of spatial data warehousing for geographic knowledge discovery,Geographic Data Mining and Knowledge discovery,Taylor&Francis,London,2001,PP.53-73.

[4] M..F.Constabile,D.Malerba(Eds),Special issue on visual data mining Journal of Visual languages and Computing14(2003)

[5] W.L.Johnston ,Model Visualization,in Information Visualization in data Mining and knowledge discovery,Morgan Kaufmann,los Altos,CA,2001,PP.223-227

