ADAPTATION AND USE OF SPATIAL AND NON-SPATIAL DATA MINING

Manuel PECH, David SOL, Jesús GONZÁLEZ
Universidad de las Américas-Puebla (UDLA)
San Andrés Cholula, Puebla, México, MEXICO
{sp205175, sol}@mail.udlap.mx, jagonzalez@inaoep.mx

ABSTRACT

The investigation described in this paper states the analysis and application of spatial and non-spatial data mining technology. Spatial data mining can be defined as the search of patterns that could exist in spatial databases. The test context is a database of the Popocatépetl volcano developed by the laboratory of Technologies of GeoInformation at UDLA-P. Spatial data mining was centered in clustering geometric objects by the implementation of the PAM (Partitioning Around Medoids) algorithm. Non-spatial data mining was made through the use of the SUBDUE system, which searches representative substructures in the data. The obtained results helped us to identified relevant characteristics which need to be improved as part of a contingency plan for the population living around the volcano risk zones (i.e. evacuation roads).

KEY WORDS

GIS, spatial data mining, non-spatial data mining, Open Gis.

1. INTRODUCTION

The Laboratory of Technologies of GeoInformation (Xaltal) [9] of the Universidad de las Americas-Puebla is developing the Popocatépetl Project [5, 11, 15, 17], which main objective is to offer a Geographic Information System (GIS) that incorporates all the information related to the Popocatépetl volcano, and allow a remote way to be accessed by different kinds of users.

A GIS is defined as a tool for geographic data manipulation [2]. It implements a great diversity of functions; some of these are the compilation, verification, storage, recovery, manipulation, updating and presentation of geographic data. One of its more important potentialities is the inclusion of modules for data analysis.

Our research involves the use of the data mining technology, applied to the volcano database, as a tool for knowledge discovery. The general objective is the adaptation, implementation and use of algorithms for spatial and non-spatial data mining to be applied to the Popocatépetl volcano database. Our intention is to provide a tool that allows the analysis, evaluation and optimization of the volcano information, as well as the advice that knowledge discovery can give at the time of decision making.

This work is divided into 7 sections. Section 1 shows the introduction to the investigation. In section 2 we explain what spatial data mining is. In section 3 we describe data mining and the SUBDUE system. In section 4 we introduce the Popocatépetl volcano database and the architecture of the developed system. Section 5 presents the implementation of the system. In section 6 some results obtained are presented and finally in section 7 the conclusions and the future work are given.

2. SPATIAL DATA MINING

Diverse studies on the methods for the spatial database knowledge discovery have been made. For example, Junas Adhikary [1] presents a classification of these methods and divides them in five groups:

1. Methods using generalization are based on knowledge discovery and require the implementation of concepts hierarchies. In the case of a spatial database, these hierarchies can be thematic or spatial. A thematic hierarchy can be exemplified when generalizing specific concepts like apples and pears to fruits; and spatial hierarchies when generalizing a group of points in a map as a region and a group of regions as a country.
2. A method using pattern recognition can be used to make automatic recognition and categorization of photographs, images, and text, among others.
3. Methods using clustering are used to create groupings or data associations, when there is some knowledge similarity among the elements of the group (i.e., similarity by Euclidian distance).
4. Methods exploring spatial associations allow discovering spatial association rules, that is, rules that associate one or more spatial objects with another or other spatial objects (X → Y, where X and Y are a set of spatial or non-spatial predicates).
5. Methods using approximation and aggregation allow knowledge discovery on the basis of the representative characteristics of the data set.

Spatial Data

Spatial data consists of information that describes spaces. Data is continuously obtained by diverse types of applications such as GIS and computerized cartography. Consequently, data analysis by means of manual techniques is sometimes a complicated task, due the large volume of data. In order to solve this problem, different methods have been proposed and applied to discover knowledge in spatial data. Most of these methods use machine learning techniques, database technology and statistics.

Spatial data mining is defined as the discovery of implicit and previously unknown knowledge in spatial databases [4]. Knowledge discovered from spatial data can be classified into several types, like representative characteristics, structures or clusters and spatial associations, just to mention a few.

Methods for knowledge discovery in spatial data

Geographic data in general has thematic and spatial data [1]. Thematic data is alphanumeric and related to the spatial objects. Spatial data, on the other hand, is described using two different properties: geometry and topology. According to [1], spatial location and size are considered geometric properties, whereas adjacency (the object A is right of object B) and inclusion (the object A is included in object B) are considered topological properties. In this way the methods discovering knowledge can be focused either on the thematic or in the spatial properties of spatial objects of a spatial database or both.

Methods using generalization

One of the most effective methods for discovering knowledge has been the learning from examples technique (with generalization). This method requires concept hierarchies.

Methods using clustering

Cluster analysis is a branch of statistics; the main advantage of using this technique is the feasibility to directly find groups (clusters) in the data without using any background knowledge, similar to an unsupervised learning approach used in machine learning. Diverse algorithms have been developed like PAM (Partitioning Around Medoids) [7], CLARA (Clustering LARge Applications) [7] and CLARANS (Clustering Large Applications based upon RANdomized Search) [12]. Next we describe the PAM algorithm.

PAM

In order to find *k* clusters (groups), PAM determines a representative object for each cluster. This representative object, called medoid, is the one that is located toward the center within the cluster. Once medoids have been selected, each non-selected object is grouped with the medoid to which is the most similar. More precisely, if O_j is a non-selected object, and O_i is a medoid (selected object), we say that O_j belongs to the cluster represented by O_i, if $d(O_j, O_i) = min_{oe}d(O_j, O_e)$, where the notation min_{oe} denotes the minimum over all medoids O_e, and the notation $d(O_a, O_b)$ denotes the dissimilarity or distance between O_a and O_b objects. All the dissimilarity values are given as inputs to PAM.

To find the *k* medoids, PAM begins with an arbitrary selection of *k* objects. Then in each step, a swap between a selected object O_i and a non-selected object O_h is made, as long as such a swap would result in an improvement of the quality of the clustering. In particular, to calculate the effect of such a swap between O_i and O_h, PAM computes cost C_{jih} cost for all non-selected objects O_j. Depending on which of the following cases O_j is in, C_{jih} is defined by one of the following equations:

Case 1

Supposing that O_j currently belongs to the cluster represented by O_i, Furthermore, letting O_j be more similar to $O_{j,2}$ than to O_h, $(d(O_j,O_h) \geq d(O_j,O_{j,2})$, where $O_{j,2}$ is the second most similar medoid to Oj. Thus, if O_i is replaced by O_h as a medoid, O_j would belong to the cluster represented by $O_{j,2}$. Hence the cost of the swap is given by $C_{jih} = d(O_j, O_{j,2})$ - d(O$_j$,O$_i$). This equation always gives a non-negative C_{jih} value, indicating that there is non-negative cost incurred in replacing O_i by O_h. An example is shown in Figure 2.1. Let us suppose that $d(O_j, O_{j,2}) = 6$ and $d(O_j, O_i) = 2$, then the C_{jih} value = 6 - 2 = 4.

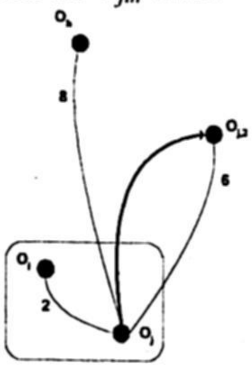

Figure 2.1. Casc 1.

Case 2

O_j currently belongs to the cluster represented by O_i. But this time, O_j is less similar to $O_{j,2}$ than O_h, $(d(O_j, O_h) < d(O_j, Oj,_2)$. Then, if O_i is replaced by O_h, O_j would belong to the cluster represented by O_h. This way the cost is given by $C_{jih} = d(O_j, O_h) - d(O_j, O_i)$. Unlike the equation in case 1, C_{jih} can be positive or negative, depending on whether O_j is more similar to O_i or to O_h.

Case 3

Supposing that O_j currently belongs to a cluster other than the one represented by O_i. In addition, assuming $O_{j,2}$ is the representative object of that cluster and O_j is more similar to $O_{j,2}$ that O_h. Then, even if O_i is replaced by O_h, O_j would stay in the cluster represented by $O_{j,2}$. Thus, the cost is given by $C_{jih} = 0$.

Case 4
O_j currently belongs to the cluster represented by $O_{j,2}$. But O_j is less similar to $O_{j,2}$ than O_h. Replacing O_i with O_h would cause O_j jump to the cluster represented by O_h from cluster $O_{j,2}$. This way the cost is given by $C_{jih} = d(O_j, O_h) - d(O_j, O_{j,2})$. This cost always is negative.

Combining the four cases, the total cost of replacing O_i with O_h is given by:

$$TC_{ih} = \sum_j C_{jih}$$

Algorithm PAM:

1. Select k representative objects arbitrarily.
2. Compute TC_{ih} for all the pairs of objects O_i,O_h where O_i is currently selected, and O_h is not.
3. Select the pair O_i,O_h which corresponds to $minO_i,O_h$ TC_{ih}. If the minimum TC_{ih} is negative, replace O_i with O_h, and go back to step 2.
4. Otherwise, for each non-selected object, find the most similar representative object.
5. Halt.

Spatial Data Mining based on Clustering Algorithms
In this section, we present two spatial data mining algorithms developed by Kaufman and Rousseeuw [7]: Spatial Dominant Approach (SD) and Non-Spatial Dominant Approach (NSD).

Spatial Dominant Approach: SD
There are different types of approaches to spatial data mining. A spatial database consists of spatial and non-spatial attributes. The non-spatial attributes are stored in relations. The general approach here is to use clustering algorithms to work with the spatial attributes, and use other learning tools to take care of non-spatial data over the spatial findings.

Non-Spatial Dominant Approach: NSD
The spatial dominant algorithms, such as SD, can be viewed as focusing asymmetrically on discovering non-spatial characterizations of spatial clusters. Non-spatial dominant algorithms, on the other hand, focus on discovering spatial clusters existing in the result of data mining in non-spatial data.

3. NON-SPATIAL DATA MINING

Data mining in general can be seen as the search for hidden patterns that may exist in databases [12]. The explosive growth in data and databases has generated a need for techniques and tools that can transform the data into useful information and knowledge.

Knowledge discovery in databases refers to the task of finding interesting knowledge, regularities, or high-level information from data sets, which can then be analyzed from different angles. Researchers in many different fields including database system, knowledge-base system, artificial intelligence, machine learning and statistic have shown great interest in data mining.

Some of the data mining techniques apply on structural data and others on non-structural data. A structural data is defined as data that describes the relationships among the objects described in the data. We can see the data objects as variables in the attribute-value representation, but now we also have relations among those variables.

In this work we used a data mining system that uses a graph-based learning technique. This technique has the potential to be competitive in the learning task, because it provides a powerful and flexible knowledge representation that can be used for relational domains.

SUBDUE.
The SUBDUE system [18] (developed at the University of Texas in Arlington) is a general tool that can be applied to any domain that can be represented as a graph. It discovers substructures that compress the original database and represents interesting structural concepts in the data. By replacing previously-discovered substructures in the data, multiple passes of SUBDUE produce a hierarchical description of the structural regularities in the data. SUBDUE has the capability to use a constrained inexact graph match that can consider similar, but not identical, instances of a substructure as a pattern. SUBDUE uses the minimum description length principle to guide the search towards more appropriate substructures.

The SUBDUE system uses a graph representation. Objects in the data (concepts) become vertices or small subgraphs in the graph, and relationships between objects become directed or undirected edges in the graph. A substructure is a connected subgraph within the graph. This graph representation serves as input to the SUBDUE system. Figure 3.1 shows an example of an input database and its graph representation. The example is presented in terms of the house domain, where a house is defined as triangle on a square. T represents a triangle, C a square, E a star and R a rectangle. The objects in the figure (T1, C1, R1) become labeled vertices in the graph, and the relationship (on, shape) become labeled edges. The graph representation of the substructure discovered by SUBDUE from this data is shown in figure 3.2.

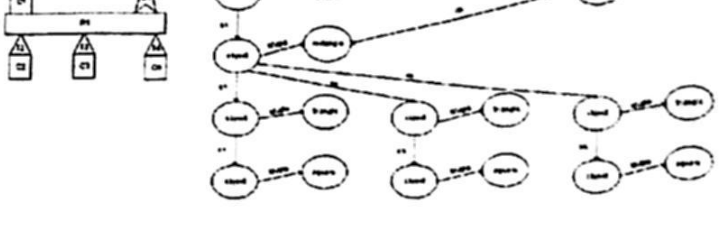

Figure 3.1. Graph Representation of the House Domain.

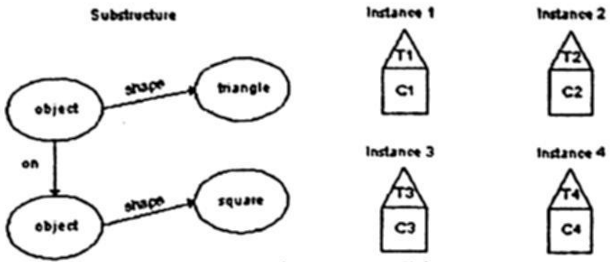

Figure 3.2. Substructure and Instances Discovered from the House Domain by Subdue.

An instance of a substructure in an input graph is a set of vertices and edges from the input graph that match the graph definition of the substructure. A neighboring edge of a substructure instance is an edge in the input graph that is not contained in the instance, but is connected to at least one vertex in the instance. An external connection of an instance of a substructure is a neighboring edge of the instance that is connected to at least one vertex not contained in the instance.

4. SPATIAL AND NONSPATIAL DATA MINING INTEGRATION

Database architecture

The database scheme complies with the Open GIS SQL92 specification [13]. The storage type used for geometries is the binary geometry schema (Well-know Binary for Representation Geometry). The database is made up of 150 entities containing descriptive and geometric information; in addition there are two more metadata catalogues.

System Architecture

The system was developed using the Java programming language and consists of 10 class packages (kdd, baseKDD, bd, shp, formats, factory, opengis, oracle, graficacion2D, paqueteDeGeometrias). Oracle was used as the database management system.

The Data Mining Process

The data mining process implemented in this project is based on the Spatial Dominant Approach (SD).

1. Selection of the data layers that will integrate our initial data set.
2. Data transformation for the application of the PAM algorithm.
3. Spatial data mining process (PAM).
4. Selection of a data subset generated by PAM (clusters).
5. Data transformation for the application of the SUBDUE system (graph creation).
6. Non-spatial data mining process (SUBDUE).
7. Pattern evaluation and interpretation.
8. Knowledge application.

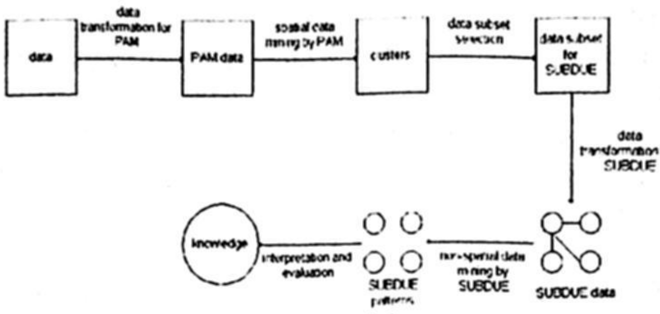

Figure 4.1. Data mining process.

5. SYSTEM IMPLEMENTATION

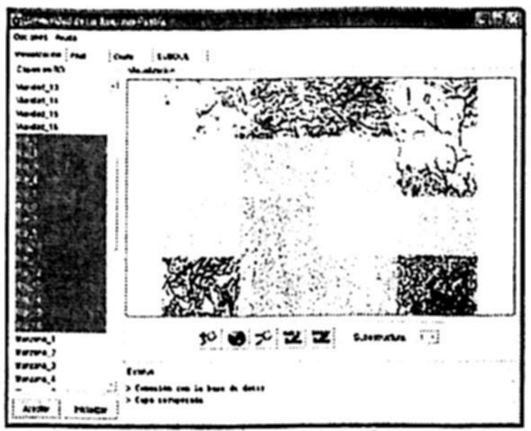

Figure 5.1. System interface.

The system is made up of the following four modules:

Visualization

This module allows the user to select and visualize the existing data layers in the database (figure 5.1). The user has the possibility to select one or more layers. Once the layers have been selected, they are displayed in the visualization area. In order to help the user to analysis them, there are options for applying colors to the lawyers for easier identification and the possibility to select a single spatial object and display its descriptive data. Additionally, we have the option to visualize the results of the data mining processes in a graphical way.

PAM

In this module the user applies the PAM data mining algorithm. Clusters of spatial objects (point, line, and polygon) are found on the basis of its closeness with other objects (using Euclidian distance). An example of an application for this technique is our wish to learn about the most important characteristics of the rivers located in the north zone of the Popocatépetl volcano. In order to delimit our data set, we can find clusters of the rivers and select the clusters within the zone of interest. Once we have identified the elements (clusters) we can apply SUBDUE to their descriptive data to find important characteristics about those rivers.

GRAPH

The Graph module is used to transform the descriptive information of the database (stored in relational form) into a graph representation. Each attribute value becomes a vertex and each attribute name becomes an arc. Once the graph has been created, it is stored in a text file. At this point, data transformation is necessary since SUBDUE requires its input data set to be in graph form.

SUBDUE
This module is used to implement non-spatial data mining. It is invoked through a call to the operating system and its results directed to a text file. Later, the text file content must be loaded into the system to be able to visualize the results graphically. In order to mining the Popocatépetl database, we use the Spatial Dominant Approach.

6. RESULTS

In this section we present an example of the results generated by the system. For the application of the PAM algorithm we used three clusters. The representation of each cluster is pointed out by the following colors: cluster 1 red, cluster 2 blue and cluster 3 green. First, we apply the PAM algorithm to find clusters from the input data, in this case the Vialidad_1 and Vialidad_2 layers. Once we found the clusters we apply SUBDUE to the associated descriptive data. The representation of the results generated by SUBDUE is pointed out by the following colors: substructure 1 red, substructure 2 blue and substructure 3 green.

Figure 6.1 shows the results generated by PAM finding three clusters from the Vialidad_1 and Vialidad_2 entities that belong to the roads layer.

Processing the descriptive information from the Vialidad_1 and Vialidad_2 entities by SUBDUE (figure 6.2), we conclude that 60.07% of all roads are of the dirt-road type (pattern). The best substructure discovered by Subdue, which is shown below, supports this conclusion (figure 6.3).

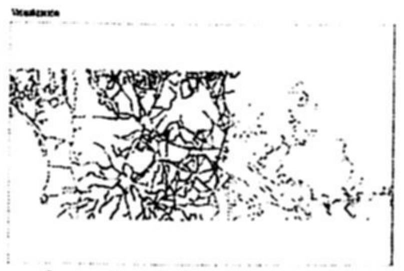

Figure 6.1. Clusters from the Vialidad_1 and Vialidad_2 entities found by PAM.

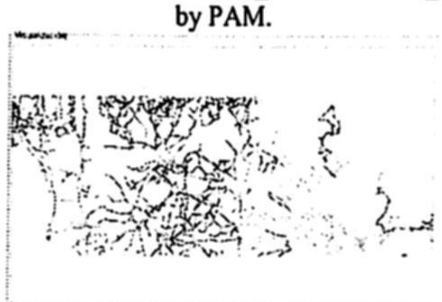

Figure 6.2. Patterns found by SUBDUE from the Vialidad_1 and Vialidad_2 layers.

Subgraph vertices
19 EVENT
24 7112.000000
25 DIRT-ROAD
26 0.000000

Subgraph edges
[19 -> 24] CODE
[19 -> 25] DESCRIPTION
[19 -> 26] ROUTE_NUMBER
Number of instances = 1345

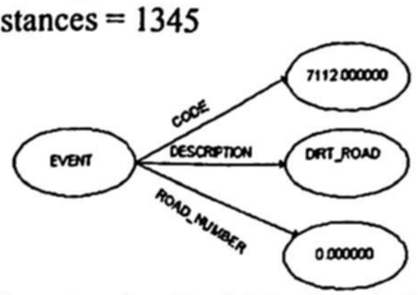

Figure 6.3. Substructure found by Subdue from the Vialidad_1 and Vialidad_2 layers.

This substructure is very important because is telling us that there are some areas that need more material roads in order to implement a contingency plan to vacate the zone. In cases like this, we can use the PAM algorithm to find smaller clusters (from areas closer to the volcano) and find interesting knowledge with the Subdue system. Now would be trying to identify clusters with high population and with no material roads so that the existing dirty-roads could become a bottle neck in case of an emergency.

7. CONCLUSIONS AND FUTURE WORK

In this project we presented a system for applying spatial and non-spatial data mining techniques to the Popocatépetl volcano database. In the first case the clustering algorithm PAM was implemented, and in the second case the substructure discovery system SUBDUE was used. The results showed an efficient yield of PAM when working with small data sets. The complexity for one iteration is $O((n-k)^2)$. The response times of SUBDUE are based on the size of the input graph and the parameters established for their operation. The system architecture is made up of modules. Each class package is designed in such away that it is possible to add, modify or replace any of its elements. This is an important feature since new data mining technologies can be implemented in order to count with a lager number of tools to allow the discovery of more useful and beneficial knowledge.

This research showed that data mining technologies, developed to be used in other research fields, are feasible to be adapted according the context of geographic data. The PAM adaptation to work with data stored in the volcano database is a good example. The results generated by SUBDUE were transformed with the purpose of showing them graphically by using geometric objects.

The system is very valuable for its use by several kinds of users (in general, for decision making) in areas like contingency planning in case of a catastrophe. In order to enlarge the capabilities of the current system we propose for further work the issues below: Research on new clustering methodology techniques.

Handling of the SUBDUE data compression process. One of the most important features of this system is the replacement in the initial data set with the discovered substructures. This process is not supported in the current system; therefore, if in the result from SUBDUE there is a substitution, this one is not reflected in the results shown in the Visualization module. Building a Data Warehouse. The next step in the Popocatépetl volcano project evolution would be the implementation of a geographic warehouse in order to work with historical data and enhance the data mining processes.

ACKNOWLEDGEMENT

National Science and Technology Council (CONACYT). Project W35804-A. Access to High Quality Digital Services and Information for Large Communities of Users.

REFERENCES

[1] Adhikary, Junas. *Knowledge Discovery in Spatial Databases, Progress and Challenges*. School of Computing Science, Simon Fraser University. 1996.

[2] Bernhardsen, Tor. *Geographic Information Systems, An Introduction*, Second Edition. John Wiley & Son Inc. 1999.

[3] Bunke, H., and G. Allermann. *Inexact graph matching for structural pattern recognition*. Pattern and Recognition Letters, 1983.

[4] Frawley, W. J., G. Piatetsky-Shapiro, and C. J. Matheus. *Knowledge Discovery in databases: An overview*.

[5] Gómez H. 2001. *Urban Analysis Tool in a GIS*. Bachelor Thesis. Computer System Engineering. Department of Computer Science, Engineering School, Universidad de las Américas-Puebla. May 2001.

[6] Holder, L. B., D. J. Cook and S. Djoko. *Substructure Discovery in the SUBDUE System*. In Proceedings of the AAAI Workshop on Knowledge Discovery in Databases. 1994.

[7] Kaufman, L., and P. J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Son Inc. 1990.

[8] Kolatch, Erica. *Clustering Algorithms for Spatial Databases: An Survey*. Department of Computer Science, University of Maryland, Collage Park. 2001.

[9] Geo-information Laboratory. Universidad de las Américas-Puebla. Internet site, visit last time October 2002. http://mailweb.udlap.mx/~gisudla/.

[10] Laurini R., and D. Thompson. *Fundamentals oj Spatial Information Systems*, Academic Press. 1992.

[11] Morales A. 2001. *Construction and Modeling of a Geographic Database*. Bachelor Thesis. Computer System Engineering. Department of Computer Science, Engineering School, Universidad de las Américas-Puebla. May 2001.

[12] Ng, Raymond T. and Jiawei Han. 1994. *Efficient and Effective Clustering Methods for Spatial Data Mining*. Proceedings of the 20th Very Large Databases Conference (VLDB 98) Santiago, Chile.

[13] Open Gis Consortium, Inc. *OpenGIS Simple Features Specification for SQL*, Revision 1.1. OpenGis Project Document 99-049. 1999. http://www.opengis.org.

[14] Piatetsky-Shapiro, G., and W. J. Frawley, editors. *Knowledge Discovery in Databases*. AAAI/MIT Press, Menlo Park, CA, 1991.

[15] Posada N. 2001. *Implementation of a CBR for a Volcano Context*. Master Thesis. Sciences in Computer System. Department of Computer Science, Engineering School, Universidad de las Américas-Puebla. December 2001.

[16] Quinlan, J. R., and R. L. Rivest. *Inferring decision trees using the minimum description length principle*. Information and Computation, 1989.

[17] Razo A. 2001. *GISELA X3: Standard Modeling oj Three-Dimensional Geometric Data with XML and Their Application in a Geographical Information System for Civil Protection*. Master Thesis. Sciences in Computer System. Department of Computer Science, Engineering School, Universidad de las Américas-Puebla. December 2001.

[18] SUBDUE. University of Texas at Arlington. Internet site, visit last time October 2002. http://cygnus.uta.edu/subdue/.

[19] Witten, Ian H., and Eibe Frank. *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers.