GEOARM: an Interoperable Framework to Improve Geographic Data Preprocessing and Spatial Association Rule Mining

Vania Bogorny, Paulo Martins Engel, Luis Otavio Alvares

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Av. Bento Gonçalves, 9500 – Campus do Vale – Bloco IV
Bairro Agronomia – Porto Alegre – RS – Brazil
{vbogorny, engel, alvares} @ inf.ufrgs.Br

Abstract *– Geographic data preprocessing is the most expensive and effort consuming step in the knowledge discovery process, but has received little attention in the literature. For the data mining step, especially for association rule mining, many different algorithms have been proposed. Their main drawback, however, is the huge amount of generated rules, most of which are well known patterns. This paper presents an interoperable framework to reduce both the number of spatial joins in geographic data preprocessing and the number of spatial association rules. Experiments showed a considerable time reduction in geographic data preprocessing and association rule mining, with a very significant reduction of the total number of rules.*

Keywords. Geographic databases, spatial association rules, framework, knowledge base, geographic data preprocessing

1. INTRODUCTION

The increasing use of geographic data in different application domains has resulted in large amounts of data stored in geographic databases (GDB). Geographic data are real world entities, also called spatial features, which have a location on Earth's surface [1]. Spatial features (e.g. Brazil, Spain) belong to a feature type (e.g. country), and have both non-spatial attributes (e.g. name, population) and spatial attributes (geographic coordinates *x,y).*

Beyond the spatial attributes, there are implicit spatial relationships, which are intrinsic to geographic data, but usually not explicitly stored in geographic databases. Because of spatial relationships, real world entities can affect the behavior of other features in the neighborhood, and due this reason they must be extracted for data mining and knowledge discovery [2]. Spatial relationships are the main characteristic which differs knowledge discovery in geographic databases from knowledge discovery in classical databases (KDD).

Figure 1 shows an example of implicit spatial relationships where gas stations and industrial residues repositories may be *close, far,* or maybe *intersect* water bodies. Considering that water analyses showed high chemical pollution, the extraction of spatial relationships among water resources, gas stations, and industrial residues repositories is of fundamental importance for knowledge discovery.

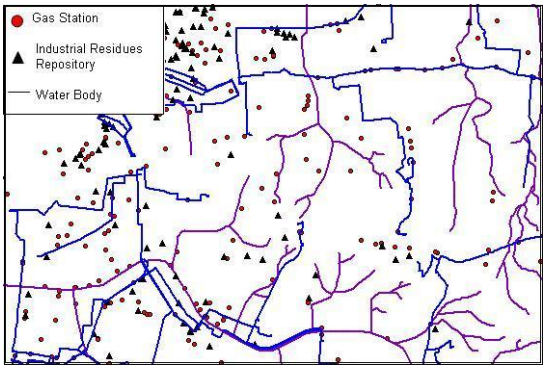

Figure 1 – Examples of implicit spatial relationships

Although spatial relationships are the main characteristic to be considered in data mining, not all relationships hold interesting information. Many relationships are well known geographic domain associations that may hind the discovery process and produce a large number of patterns without novel and useful knowledge.

Figure 2 shows an example of well known geographic domain dependences between gas stations and streets. Notice that there is an explicit pattern: all gas stations intersect streets. Such relationships produce patterns with 100% confidence in the discovery process.

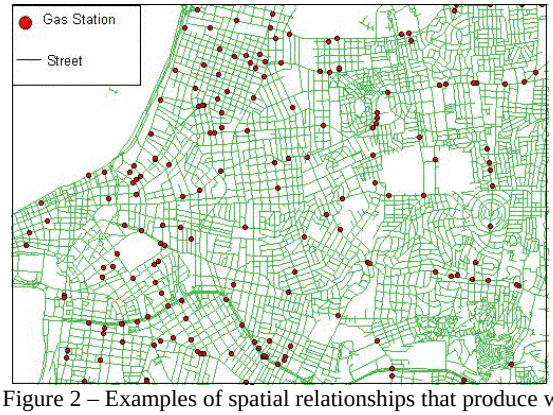

Figure 2 – Examples of spatial relationships that produce well known geographic domain patterns

In spatial association rule mining, which is a data mining technique that has been extensively used to extract knowledge from GDB, well known associations

(dependences) generate two main problems:

a) *Data preprocessing:* a large computational time is required to preprocess GDB to extract spatial relationships. The spatial join (cartesian product) operation, required to extract spatial predicates, is the most expensive operation in databases and the big processing bottleneck of spatial data analysis and knowledge discovery;

b) *Association rule mining:* a large number of spatial association rules without novel, useful, and interesting knowledge is generated.

Although users may be interested in high confidence rules, not all strong rules necessarily hold considerable information. Moreover, the mixed presentation of hundreds or thousands of interesting and uninteresting rules can discourage users from interpreting them all in order to find 'patterns' of novel and unexpected knowledge.

1.1 The Problem of Mining Association Rules with Well Known Dependences

We illustrate the problem of mining spatial association rules with well known geographic associations through a small experiment. Every row of the dataset mined was a district of the city of Porto Alegre – Brazil (target feature type), and the columns were two non-spatial attributes (hepatitis rate and sanitary condition) and three relevant spatial feature types with a spatial relationship with district (water bodies, slums, and treated water network). We wanted to investigate associations to high hepatitis incidence. District and treated water network have a well known geographic dependence because every water network belongs to one or more districts and all districts have water networks. We mined spatial association rules with three different values of minimum support, and analyzed every generated rule. The summary of the results is shown in Table 1.

Table 1. Spatial association rule reduction by eliminating one dependence

Min. Sup.	Mining With Dependences			Removing dependence in data pre-processing
	Total Numb. rules	Rules with the dependence	Rules without the dependence	
0.1	203	163	40	40
0.15	91	78	13	13
0.2	32	31	1	1

A total of 203 rules was generated for minimum support 0.1. Among the 203 rules, 163 had the dependence, and only 40 were generated without the dependence. In this case the user would have to analyze 203 rules, among which only 40 would be the most interesting, because 163 had the well known dependence.

Notice that even increasing minimum support to 0.2, from a total number of 32 rules, 31 were still generated with the dependence. This shows that minimum support does not warrant the elimination of rules with well known associations, and increasing minimum support may eliminate interesting rules.

By removing the dependence in data preprocessing (column on the right in Table 1), association rule mining algorithms will generate only rules without the dependence.

Existing algorithms for mining spatial association rules [3][4][5][6] are not intelligent enough to decide which spatial relationships are relevant to the discovery process and which are well known geographic domain associations. Geographic domain associations represent spatial integrity constraints that are explicitly represented in geographic database schemas and geo-ontologies. However, such knowledge repositories have not been used to improve the KDD process.

In this paper we present an intelligent framework to preprocess geographic databases and eliminate well known dependences for mining spatial association rules.

1.2 Related Works and Contribution

Existing approaches for mining spatial association rules do neither make use of prior knowledge to specify which spatial relationships should be computed in data preprocessing nor to reduce the number of well known patterns. Koperski [3] presented an approach for mining spatial association rules which is a top-down, progressive refinement method. In a first step spatial approximations are calculated, and in a second step, more precise spatial relationships are computed to the result of the first step. Minimun support is used in data preprocessing to extract only frequent spatial relationships. The method has been implemented in the module Geo-Associator of the GeoMiner [7] system. A similar method has been proposed by [5] for geographic objects with broad boundaries.

Appice [4] proposed an upgrade of Geo-Associator to first-order logic, and all spatial features and spatial relationships are extracted from spatial databases and represented on a deductive relational database. This process is computationally expensive and non-trivial with real data.

Mennis [6] applied Apriori[8] to geographic data to extract spatial association rules. Data preprocessing is performed with the operations provided by GIS, but no prior knowledge is used to either improve data preprocessing or prune well known association rules.

In these approaches rules are filtered by thresholds of minimum support and minimum confidence, which do not warrant the elimination of well known geographic dependences. The main difference from these approaches and our framework is that we automatically preprocess geographic databases using prior knowledge to reduce both the number of spatial joins and the number of rules.

The main contribution of this paper is the reduction of well known geographic patterns and the reduction of the computational time to preprocess geographic databases and the extraction of association rules.

The proposed framework was implemented in Weka [9], which we extended to support dynamic geographic data preprocessing, discretization, and transformation for mining spatial association rules at different granularity levels. Weka is a well established free and open source toolkit with friendly and graphical user interface which covers the whole KDD process. Other systems such as ARES or GeoMiner do not provide the same advantages, or are no longer available outside academic institutions. Moreover, any association rule mining algorithm may be applied using our framework, since the well known associations are pruned in data preprocessing steps.

1.3 Scope and Outline

The proposed framework is and extension of [10] to improve the KDD process using prior knowledge. The focus in this paper is to show how prior knowledge can be used to improve geographic data preprocessing and spatial association rule mining, and not how to represent geographic knowledge into a knowledge base.

The remainder of the paper is organized as follows: Section 2 introduces geographic domain associations and shows how they are represented in geographic database schemas and geo-ontologies. Section 3 presents the framework for enhancing geographic data preprocessing and spatial association rule mining. Section 4 shows experiments with real databases, while Section 5 concludes the paper and suggests some directions of future work.

2. GEOGRAPHIC DOMAIN ASSOCIATIONS

Well known geographic associations are mandatory spatial relationships, usually represented as spatial integrity constraints [11] in geographic database schemas [12][13] and geo-ontologies [14]. They can also be provided by the user, and in this case, a larger set of associations can be specified; not only associations explicitly represented in the schema, but many other geographic domain associations which produce well known patterns.

In geographic database schemas, well known geographic associations are normally represented by *one-one* and *one-many* cardinality constraints [13] that must hold in order to warrant the consistency of the data. Figure 3 shows an example of a geographic database schema, represented in a UML [15] class diagram. Explicit mandatory *one-one* and *one-many* relationships between gas stations and streets, county and streets, water resources and counties, as well as islands and water resources will produce well known rules because they always hold if the database is consistent. Notice that gas stations and water resources or islands and gas stations do not have any *explicit* mandatory relationship represented in the schema, so their *implicit* relationships may be interesting for knowledge discovery.

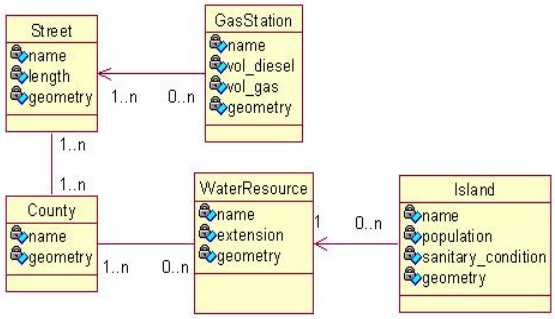

Figure 3 - Part of a conceptual geographic database schema

In the logical level, *one-one* and *one-many* relationships normally result in foreign-keys in relational geographic databases, and in pointers to classes in object-oriented geographic databases [13]. So they can be automatically retrieved with processes of reverse engineering [16] if the schema is not available.

Many different approaches to extract associations from relational databases using reverse engineering are available in the literature. For data mining and knowledge discovery in non-geographic databases reverse engineering has been used to understand the data model in legacy systems [17], or to automatically extract SQL queries [18], but not as prior knowledge to reduce well known patterns.

Ontologies are also rich knowledge repositories that have been used recently in many and different fields in Computer Science. Although research is not so far yet in ontologies for geographic data, some geo-ontologies have been emerging recently. Besides defining a geo-ontology for administrative data for the country of Portugal, Chaves [19] defined a geo-ontology meta-model, named GKB (Geographic Knowledge Base). In [14] we extended this approach to support spatial integrity constraints.

A mandatory *one-one* relationship between island and water resource, for example, may be specified in a geo-ontology with cardinality constraints, as shown bellow.

```
<owl:Class rdf:ID="Island">
  <rdfs:SubClassOf rdf:resource="#SpatialFeatureType"/>
  <rdfs:subClassOf>
   <owl:Restriction>
    <owl:minCardinality
     rdf:datatype="http://www.w3.org/2001/XMLSchema#int">
     1</owl:minCardinality>
    <owl:allValuesFrom rdf:resource="#WaterResource"/>
    <owl:OnProperty>
     <owl:ObjectProperty rdf:about="#Within"/>
    </owl:OnProperty>
   </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

Geographic associations produce well known patterns independently of the type of topological (e.g. *touches*, *contains*, *crosses*, etc) or distance relationship. For example, if there is a gas station, then there must be a street,

independently if the topological relationship is *touches, crosses, within* or *intersects.* A large number of non-novel rules is generated for each different topological relationship occurring more often than a specified minimum support (e.g. *contains (GasStation)→* ***crosses****(Street); contains (GasStation)→* ***touches****(Street)).* Due this reason, if a pair of geographic feature types is specified in the knowledge base as a pair with a well known dependence, then no spatial relationship among the pair needs to be computed.

3. THE *GEOARM* FRAMEWORK

In order to optimize GDB preprocessing and spatial association rule mining, Figure 4 shows an interoperable framework with support to the whole discovery process using the association rule mining technique. It is composed of three abstraction levels, as described in [10]: data mining, data preparation, and data repository.

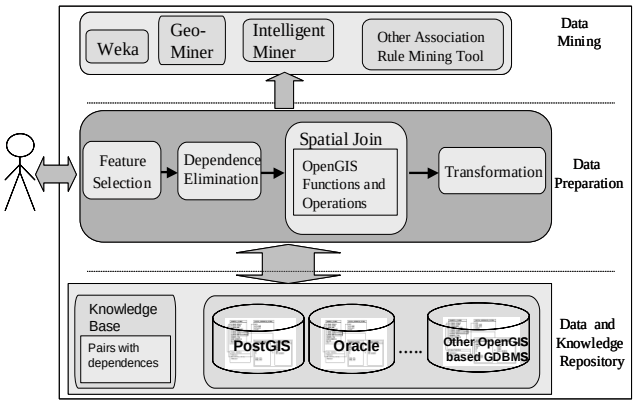

Figure 4 – GEOARM (Geographic Association Rule Miner)

At the bottom are the geographic data repositories, stored in GDBMS (geographic database management systems), constructed under OGC specifications. There is also a knowledge repository which stores well known geographic associations, extracted from database schemas, geo-ontologies, or provided by the user. At the top are the data mining toolkits or algorithms for mining association rules. In the center is the spatial data preparation level which covers the *gap* between data mining tools and geographic databases. In this level the data repositories are accessed through JDBC connections and data are retrieved, preprocessed, and transformed into the single table format with SQL statements, according to the user specifications.

There are four main modules to implement the tasks of geographic data preparation for association rule mining: *feature selection, dependence elimination, spatial join,* and *transformation,* which are described in the sequence.

3.1 Feature Selection and Dependence Elimination

The *Feature Selection* module retrieves all relevant information from the database, including the target feature type (TFT), the target feature non-spatial attributes and the set of relevant feature types (RFT) that may have some influence on the TFT. The feature types are retrieved through the OpenGIS database schema, stored in the table GEOMETRY_COLUMNS.

The *Dependence Elimination* module verifies all associations between the target feature type and all relevant feature types. It searches the knowledge base and if the TFT has a dependence with a RFT, then the RFT is eliminated from the set *S* of relevant feature types. Notice that for each relevant feature type removed from the set *S*, no spatial join is required to extract spatial relationships. By consequence, no spatial association rule will be generated with this relevant feature type.

3.2 Spatial Join

The *Spatial Join* module computes and materializes the user-specified spatial relationships between the TFT and the RFT, retrieved by the *Feature Selection* module and filtered by *Dependence Elimination* module. Three types of spatial relationships are materialized:

a) *Topological*: computes the detailed topological relationships (e.g. touches, contains);
b) *Intersection*: extracts more general topological relationships, i.e., if the TFT intersects or not the RFT;
c) *Distance*: extracts *close* and *far* distance relations. *Close* is dominant over *far* in the feature type granularity level due the fact that close objects are more co-related than objects that are far. For instance, if an instance of the TFT is close to some instances of a RFT and far from others, *close* is materialized and *far* is unconsidered.

Spatial joins to extract spatial predicates are performed on-the-fly with operations provided by the GDBMS, and only over the relevant feature types defined by the user. We follow the Open GIS Consortium (OGC) [1] specifications, what makes GEOARM interoperable with all GDBMS constructed under OGC specifications (e.g. Oracle, PostGIS, MySQL, etc). The OGC is not-for-profit organization dedicated to provide interoperability for Geographic Information Systems. Besides a standard set of operations to compute spatial relations for SQL, implemented by most GDBMS, the OGC also defines a database schema for storage of spatial data with all database characteristics.

Before compute spatial joins, MBR (minimum boundary rectangle) is performed for accelerating the extraction of spatial relationships. The *Spatial Join* module output is stored into a temporary database table *T* with the following attributes: *TFT* instance identifier, the spatial relation between *TFT* and *RFT*, the *RFT* name, and the *RFT* identifier.

3.3 Transformation

The *Transformation* module transposes as well as discretisizes the *Spatial Join* module output (table *T*) into the single table representation, understandable by

association rule mining algorithms. This step is based on SQL statements, where a single table *ST* is created with the TFT non-spatial attributes and all its instances. Then for each instance in *ST*, a new attribute is created for every different RFT in *T*, which has a spatial relation with the target feature.

This module requires two user-specified parameters: relevant features granularity - defines the level of data representation for data mining, which can be on feature instance (e.g. factory A, factory B) or feature type (e.g. factory) level; - data mining algorithm - defines the exact output format which can lithely vary according to each rule mining algorithm.

4. EXPERIMENTS AND EVALUATION

In order to evaluate the interoperability of the framework, experiments were performed with real geographic databases stored under Oracle 10g and PostGIS. Census sectors, a database table with 2157 polygons and non-spatial attributes such as population, sanitary condition, etc, was defined as the TFT. Datasets with different relevant feature types (e.g. bus routes – 4062 multilines, slums – 513 polygons, water resources – 1030 multilines) were preprocessed and mined, using prior knowledge and without using prior knowledge. The process was performed with the extended Weka and the Apriori algorithm, using different values of minimum support.

4.1 Evaluation of the Method for Data Preprocessing

Considering the granularity level of feature type and two dependences among the target feature type and the relevant feature types, Figure 5 shows a graph with the computational time reduction using our framework for data preprocessing.

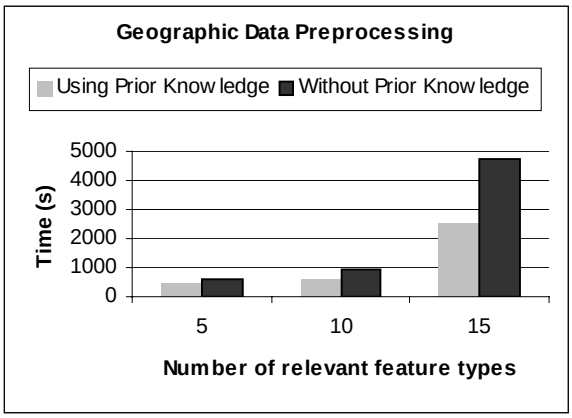

Figure 5 - Geographic data preprocessing

As can be observed in Figure 5, time reduction when preprocessing geographic databases using prior knowledge is significant even when only two well known dependences are eliminated. However, time decreases according to both number of instances and geometry type of the RFT being eliminated. For example, being the instances of the eliminated RFT 10.000 *polygons*, the time reduction would be much more significant than if the instances were, for example, 2.000 *points*.

In summary, it is difficult to precise the time reduction in data preprocessing, which varies according to the data. However, the elimination of relevant feature types avoids the spatial join operation, and this warrants the time reduction in data preprocessing.

4.2 Evaluating the Method for Mining Association Rules

Figure 6 describes one of the association rule mining experiments, with the elimination of 2 well known dependences. Notice that our framework eliminated around 68% of the number of rules even when only one dependence is removed.

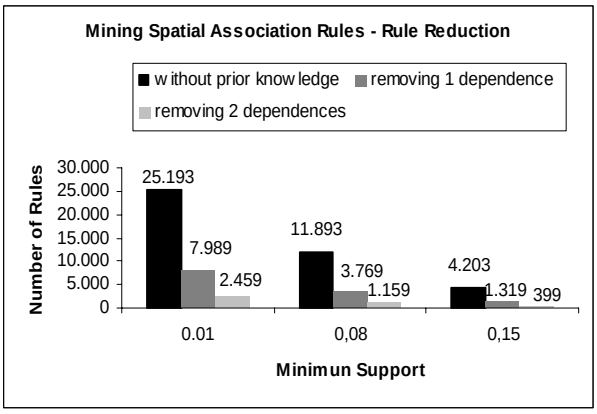

Figure 6 - Association rule reduction using Apriori

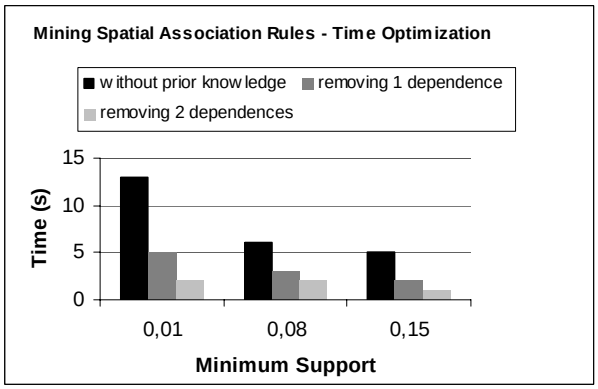

Figure 7- Computational time reduction using Apriori

When 2 dependences are removed, our framework eliminated 90% of the rules. While other methods prune rules using higher minimum support, our framework reduces rules independently of this threshold. Moreover, increasing minimum support can eliminate interesting rules, while our method warrants that only rules with well known dependences will be eliminated.

Figure 7 shows the time reduction in association rule mining when prior knowledge is used in data preprocessing.

Notice that for any value of minimum support, the time reduction remains constant, reaching 50% even if only one dependence is removed.

5. CONCLUSIONS AND FUTURE WORK

In this paper we presented a framework for mining spatial association rules from geographic databases using prior knowledge in data preprocessing. The domain knowledge refers to mandatory geographic associations which are explicitly represented in geographic database schemas and geo-ontologies. We showed that explicit mandatory relationships produce irrelevant patterns, while the implicit spatial relationships may lead to more interesting rules.

Experiments showed that independent of the number of elements, one dependence is enough to prune a large number of rules, and the higher the number of the well known dependences, the larger is the rule reduction.

The main contribution of our approach is for the data mining user, which will to analyze much less obvious rules. The method is effective independently of other thresholds, and it warrants that geographic domain associations between the target feature type and the relevant feature types will not appear among the set of rules.

The use of prior knowledge in geographic data preprocessing has three main advantages: spatial relationships between features with dependences are not computed; time reduction is very significant in data preprocessing and rule mining; and the most significant, the reduction of the number of rules.

One limitation of the framework, however, is that only well known associations between the target feature type and the relevant feature types are eliminated. As future work we will remove well known associations among relevant feature types, which can only be performed into the association rule mining algorithm.

ACKNOWLEDGMENT

Our thanks for both CAPES and CNPQ which partially provided the financial support for this research. To Procempa, for the geographic database.

REFERENCES

1. Open Gis Consortium. Topic 5, the *OpenGIS* abstract specification – *OpenGIS* features – Version 4 (1999). Available at: http://www.*OpenGIS*.org/techno/specs.htm. Accessed in August (2005).
2. Ester, M., Kriegel, H-P, Sander, J. (1997). Spatial Data Mining: A Database Approach. In Proceedings of the 5th International Symposium on Large Spatial Databases (SSD'07), Springer, Berlin, Germany, pp. 47-66.
3. Koperski, K., and Han, J. Discovery of spatial association rules in geographic information databases. In Proceedings of the SSD 4th international Symposium in large Spatial Databases, (SSD'95) Springer, Portland, Maine, USA, 1995, p.47-66.
4. Appice, A., et al. Discovery of spatial association rules in geo-referenced census data: A relational Mining Approach. Intelligent Data Analysis. Vol. 6, 2003.
5. Clementini, E., Di Felice, P., Koperski, K. Mining multiple-level spatial association rules for objects with a broad boundary. *Data & Knowledge Engineering*, 34(3):251-270 (2000)
6. Mennis, J., Liu, J.W. Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and land cover change. Transactions in GIS 9 (1), 2005, 5-17.
7. Han, J., Koperski, K., Stefanovic, N. GeoMiner: A System Prototype for Spatial Data Mining, in SIGMOD 1997, J. Peckham, ed., Proceedings of the ACM-SIGMOD International Conference on Management of Data, *SIGMOD Record* **26**(2) (1997), 553–556.
8. Agrawal, R., Srikant, R. Fast algorithms for mining association rules. In Proceedings of the 20th Int'l Conference on Very Large Databases (VLDB'94). (September 1994), Santiago, Chile, 1994.
9. Witten, I., H., Frank, E. Data Mining: Practical Machine Learning Tools And Techniques With Java Implementations. Morgan Kaufmann Publishers, San Francisco, CA, 2000.
10. Bogorny, V., Engel, P. M., and Alvares, L.O. A reuse-based spatial data preparation framework for data mining. In Proceedings of the 17th International Conference on Software Engineering and Knowledge Engineering, (SEKE'05). Jul. 2005, Taipei, Taiwan, 2005, 649-652.
11. Servigne, S. et al. A Methodology for Spatial Consistency Improvement of Geographic Databases. *Geoinformatica*, Kluwer Academic Publishers, Hingham 4-1 (2000) 7-34.
12. Bogorny, V., Iochpe, C. Extending the OpenGIS Model to Support Topological Integrity Constraints. In *Proceedings of the Brazilian Symposium on Databases*, (SBBD'2001) COPPE/UFRJ, Rio de Janeiro, Brazil, 2001, 25-39 .
13. Shekhar, S., and Chawla, S. Spatial databases: a tour. Prentice Hall, Upper Saddle River, NJ, 2003.
14. Bogorny, V., Engel, P. M., Alvares, L.O. Towards the Reduction of Spatial Join for Knowledge Discovery in Geographic Databases using Geo-Ontologies and Spatial Integrity Constraints. In: Second Workshop on Knowledge Discovery and Ontologies (KDO'2005), in conjunction with the ECML/PKDD, Porto, Portugal, 2005, 51-58.
15. Booch, G., Rumbaugh, J. and Jacobson, I. The unified modeling language: user guide. Addison-Wesley, 1998.
16. Chifosky, E.J., Cross, J.H. Reverse Engineering and Design Recovery: a taxonomy. IEEE Software, Jan .1990.
17. MCKearney, S., Roberts, H. Reverse engineering databases for knowledge discovery. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). Portland, Oregon, August 1996. 375-378.
18. Shoval, P., Shreiber, N. Database reverse engineering: from the relational to the binary relationship model. Data and Knowledge Engineering. 10: 293-315, 1993.
19. Chaves, M. S. et al. A Geographic Knowledge Base for SemanticWeb Applications. In: Brazilian Symposium on Databases (SBBD'2005), Uberlandia, Brazil (2005).