Chapter IX
Enhancing the Process of Knowledge Discovery in Geographic Databases Using Geo-Ontologies

Vania Bogorny
Universidade Federal do Rio Grande do Sul (UFRGS), Brazil

Paulo Martins Engel
Universidade Federal do Rio Grande do Sul (UFRGS), Brazil

Luis Otavio Alavares
Universidade Federal do Rio Grande do Sul (UFRGS), Brazil

ABSTRACT

This chapter introduces the problem of mining frequent geographic patterns and spatial association rules from geographic databases. In the geographic domain most discovered patterns are trivial, non-novel, and noninteresting, which simply represent natural geographic associations intrinsic to geographic data. A large amount of natural geographic associations are explicitly represented in geographic database schemas and geo-ontologies, which have not been used so far in frequent geographic pattern mining. Therefore, this chapter presents a novel approach to extract patterns from geographic databases using geo-ontologies as prior knowledge. The main goal of this chapter is to show how the large amount of knowledge represented in geo-ontologies can be used to avoid the extraction of patterns that are previously known as noninteresting.

INTRODUCTION

Knowledge discovery in databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from data (Fayyad et al., 1996). In frequent pattern mining (FPM), which is the essential role in mining associations, one of the main problems is the large amount of generated patterns and rules. In geographic databases this problem increases significantly because most discovered patterns include well-known natural associations intrinsic to geographic data. While in transactional databases items are supposed to be independent from each other (e.g.,, milk, cereal, bread), independently of their meaning, in geographic databases a large amount of data are semantically dependent (e.g., island ***within*** water).

Geographic dependences are semantic constraints that must hold in geographic databases (GDB) to warrant the consistency of the data (e.g.,, island must be completely located inside a water body). They are part of the concept of geographic data and are explicitly represented in geo-ontologies. Without considering semantics of geographic data, the same geographic dependences explicitly represented in geo-ontologies and geographic database schemas are unnecessarily extracted by association rule mining algorithms and presented to the user.

Geographic dependences produce two main problems in the process of mining spatial association rules:

a. **Data preprocessing:** A large computational time is required to preprocess GDB to extract spatial relationships (e.g., ***intersection*** between districts and water bodies). The spatial join (Cartesian product) operation, required to extract spatial relationships, is the most expensive operation in databases and the processing bottleneck of spatial data analysis and knowledge discovery.

b. **Frequent pattern and association rule generation:** A large number of patterns and spatial association rules without novel, useful, and interesting knowledge is generated (e.g., ***is_a(Island) → within (Water)***).

Aiming to improve geographic data preprocessing and eliminate well-known geographic dependences in geographic FPM in order to generate more interesting spatial association rules (SAR), this chapter presents a unified framework for FPM considering the semantics of geographic data, using geo-ontologies. While dozens of spatial and nonspatial FPM algorithms define syntactic constraints and different thresholds to reduce the number of patterns and association rules, we consider ***semantic knowledge constraints*** (Bogorny et al., 2005b)**,** and eliminate the exact sets of geographic objects that produce well-known patterns (Bogorny et al., 2006b, 2006c).

The main objective of this chapter is to show the important role that ontologies can play in the knowledge discovery process using the FPM technique. The focus addresses the use of semantic knowledge stored in ontologies to reduce uninteresting patterns, but not to create ontologies for data mining.

The remainder of the chapter is organized as follows: Section 2 presents some background concepts about geographic data, spatial relationships, spatial integrity constraints, and geo-ontologies. Section 3 introduces the concepts of frequent patterns and spatial association rules, the problem generated by geographic dependences in both data preprocessing and spatial association rule mining, and what has been done so far to alleviate this problem. Section 4 presents a framework to improve geographic data preprocessing and spatial association rule mining using geo-ontologies. Experiments are presented to show the significant reduction in the number of frequent patterns and association rules. Section 5 presents future trends and Section 6 concludes the chapter.

BACKGROUND

Geographic data are real world entities, also called spatial features, which have a location on Earth's surface (Open GIS Consortium, 1999a). Spatial features (e.g., Brazil, Argentina) belong to a feature type (e.g., country), and have both nonspatial attributes (e.g., name, population) and spatial attributes (geographic coordinates ***x,y*).** The latter normally represent points, lines, polygons, or complex geometries.

In geographic databases, every different feature type is normally stored in a different database relation, since most geographic databases follow the relational approach (Shekhar & Chawla, 2003). Figure 1 shows an example of how geographic data can be stored in relational databases. There is a different relation for every different geographic object type (Shekhar & Chawla, 2003) street, water resource, and gas station, which can also be called as ***spatial layers.***

The spatial attributes of geographic object types, represented by ***shape*** in Figure 1, have implicitly encoded spatial relationships (e.g., close, far, contains, intersects). Because of these relationships, real world entities can affect the behavior of other features in the neighborhood. This makes spatial relationships the main characteristic of geographic data to be considered for data mining, knowledge discovery (Ester et al., 2000; Lu et al., 1993), and the main characteristic, which separates spatial data mining from nonspatial data mining.

The process of extracting spatial relationships brings together many interesting and uninteresting spatial associations. Figure 2 shows an example where gas stations and industrial residues repositories may have any type of spatial relationship with water resources. Considering, for example, that water analysis showed high chemical pollution, the different spatial relationships among water resources, gas stations, and industrial residues repositories will be interesting for knowledge discovery. Notice in Figure 2 that there is a standard pattern among the data.

Figure 3 shows two examples of spatial relationships that represent well-known geographic domain dependences. In Figure 3 (left), viaducts intersect streets, and bridges intersect both water resources and streets, since both bridges and viaducts have the semantics of connecting streets. In Figure 3 (right), gas stations intersect streets they do only exist in areas with streets access.

Figure 1. Example of geographic data storage in relational databases

(a) Street

Gid	Name	Shape
1	Erie ST	Multiline [$(x_1,y_1),(x_2,y_2),..$]
2	Oak ST	Multiline [$(x_1,y_1),(x_2,y_2),..$]

(b) WaterResource

Gid	Name	Shape
1	Jacui	Multiline [$(x_1,y_1),(x_2,y_2),..$]
2	Guaiba	Multiline [$(x_1,y_1),(x_2,y_2),..$]
3	Uruguai	Multiline [$(x_1,y_1),(x_2,y_2),..$]

(c) GasStation

Gid	Name	VolDiesel	VolGas	Shape
1	BR	20000	85000	Point[(x_1,y_1)]
2	IPF	30000	95000	Point[(x_1,y_1)]
3	Elf	25000	120000	Point[(x_1,y_1)]

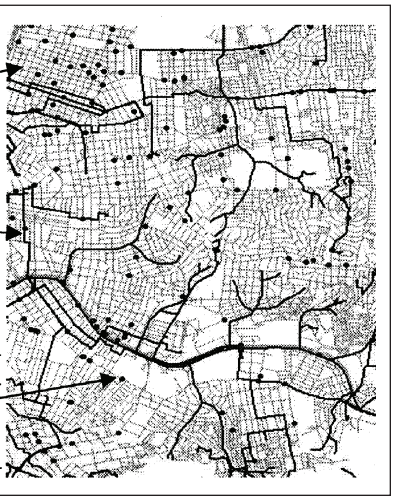


Figure 2. Examples of implicit spatial relationships

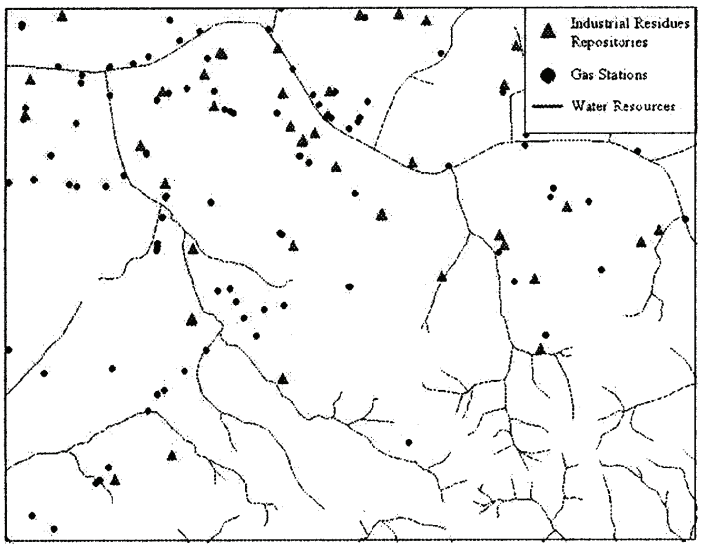
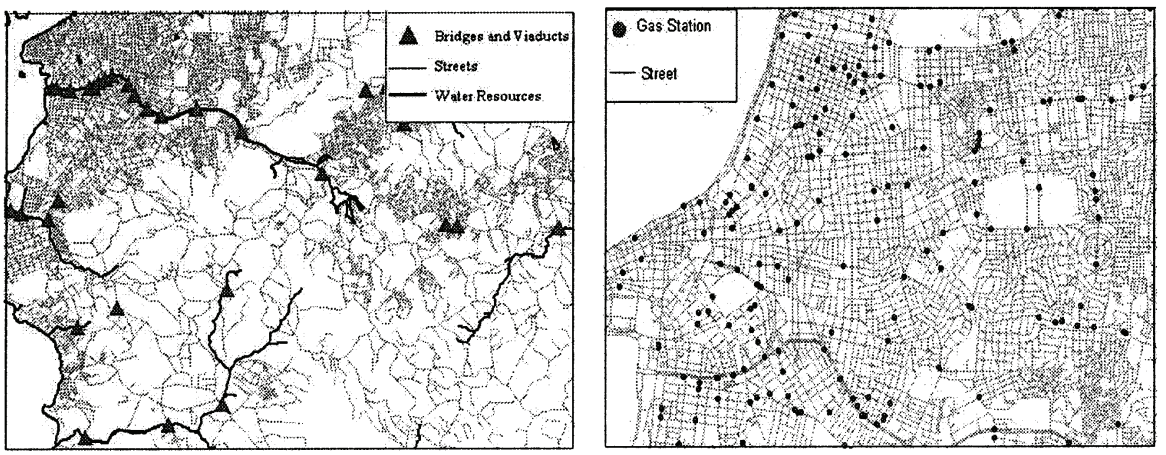


Figure 3. Examples of spatial relationships that produce well known geographic patterns in spatial data mining



The main difference between the examples shown in Figure 2 and Figure 3 is that in the former spatial relationships may hold or not, and may conduce to more interesting patterns. In the latter, under rare exceptions or some geographic location inconsistency, the spatial relationships hold for practical purposes in a 100% of the cases, and will produce well known geographic domain patterns in the discovery process. If considered in association rule mining, well known spatial relationships will generate high confidence rules such as ***is_a*** (***Viaduct***) → ***intersect*** (***Street***) (99%) or ***is_a*** (***GasStation***) → ***intersect*** (***Street***) (100%). Although users might be interested in high confidence rules, not all strong **rules** necessarily hold considerable information. Moreover, the mixed presentation of thousands of interesting and uninteresting rules can discourage users

from interpreting them in order to find novel and unexpected knowledge (Appice et al., 2005).

Patterns in the discovery process should be considered interesting when they represent ***unknown*** strong regularities, rare exceptions, or when they help to distinguish different groups of data. In geographic databases, however, there are a large number of patterns intrinsic to the data, which represent strong regularities, but do not add novel and useful knowledge to the discovery. They are mandatory spatial relationships which represent spatial integrity constraints that must hold in order to warrant the consistency of geographic data.

Spatial Relationships and Spatial Integrity Constraints

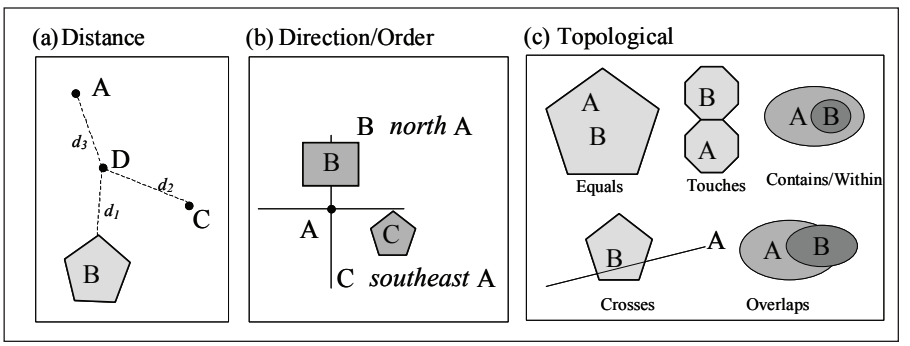
Spatial relationships can be classified as ***distance, direction,*** and ***topological***. ***Distance*** relationships are based on the Euclidean distance between two spatial features, as shown in Figure 4(a). ***Direction*** relationships deal with the order as spatial features are located in space such as north, south, east, and so forth, as shown in Figure 4(b). ***Topological*** relationships describe concepts of adjacency, containment, and intersection between two spatial features, and remain invariant under topological transformations such as rotating and scaling. Figure 4(c) shows examples of topological relationships, which will be the focus in this chapter.

Binary topological relationships are mutually exclusive, and there are many approaches in the literature to formally define a set of topological relationships among points, lines, and polygons (Clementini et al., 1993; Egenhofer & Franzosa, 1995). The OGC (Open GIS Consortium) (Open GIS Consortium, 2001), which is an organization dedicated to develop standards for spatial operations and spatial data interchange to provide interoperability between Geographic Information Systems (GIS), defines a standard set of topological operations: ***disjoint, overlaps, touches, contains, within, crosses,*** and ***equals***.

Topological relationships can be ***mandatory, prohibited,*** or ***possible***. Mandatory and prohibited spatial relationships represent spatial integrity constraints (Cockcroft, 1997; Serviane et al., 2000), and their purpose is to warrant as well as maintain both the quality and the consistency of spatial features in geographic databases.

Mandatory spatial integrity constraints are normally represented by cardinalities ***one-one*** and ***one-many*** in geographic data conceptual modeling (Bogorny et al., 2001; Serviane et al., 2000; Shekhar & Chawla, 2003) in order to warrant that every instance of a geographic feature type is spatially related to at least one instance of another spatial feature type (e.g., "island within water body"). In data mining, such constraints produce well-known patterns and high confidence rules because of the strong co-relation of the data.

Figure 4. Spatial relationships



While mandatory relationships must hold, ***prohibited*** relationships should not (e.g., "road cannot contain river").

Possible relationships, however, are usually not explicitly represented, since they can either exist or not (e.g., "roads cross water bodies," "counties contain factories"). ***Possible*** relationships may produce more interesting patterns, and are therefore the most relevant to find novel and useful knowledge in spatial data mining.

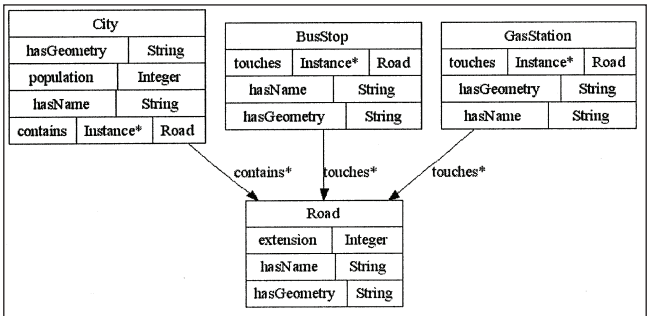
Mandatory constraints are well-known concepts to geographers and geographic database designers, and are normally explicitly represented in geographic database schemas (Bogorny et al., 2006b, 2006c) and geo-ontologies (Bogorny et al., 2005b).

Geo-Ontologies and Spatial Integrity Constraints

Ontology is an explicit specification of a conceptualization (Gruber, 1993). More specifically ontology is a logic theory corresponding to the intentional meaning of a formal vocabulary, that is, an ontological commitment with a specific conceptualization of the world (Guarino, 1998). It is an agreement about the concepts meaning and structure for a specific domain. Each concept definition must be unique, clear, complete, and nonambiguous. The structure represents the properties of the concept, including a description, attributes, and relationships with other concepts.

Ontologies have been used recently in many

Figure 5. Geo-Ontology representation and OWL code




```
<owl:Class rdf:ID="City">
 <rdfs:subClassOf>
  <owl:Restriction>
   <owl:valuesFrom rdf:resource="#Road"/>
   <owl:minCardinality rdf:datatype="#int" >1</owl:minCardinality>
   <owl:onProperty>
    <owl:ObjectProperty rdf:ID="contains"/>
   </owl:onProperty>
  </owl:Restriction>
 </rdfs:subClassOf>
 <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
</owl:Class>
<owl:Class rdf:ID="GasStation">
 <rdfs:subClassOf>
  <owl:Restriction>
   <owl:minCardinality rdf:datatype="#int" >1</owl:minCardinality>
   <owl:onProperty>
    <owl:ObjectProperty rdf:about="#touches"/>
   </owl:onProperty>
   <owl:valuesFrom rdf:resource="#Road"/>
  </owl:Restriction>
 </rdfs:subClassOf>
 <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
</owl:Class>
```

and different fields in computer science, such as artificial intelligence, databases, conceptual modeling, semantics Web, and so forth. Therefore, a relevant number of ontologies has been proposed, and a number of models, languages, and tools was developed. Chaves et al. (2005a), besides defining a geo-ontology for administrative data for the country of Portugal, defines a metamodel, named GKB (geographic knowledge base), which is a starting point to define an ontology for geographic data.

In geo-ontologies, spatial integrity constraints are represented by properties of geographic data. They are specified as restriction properties given by a spatial relationship and both minimum and maximum cardinalities. For instance, a concept ***island***, which is a piece of land surrounded by water, must have a mandatory ***one-one*** relationship with the concept ***water.***

Figure 5 shows a small example of a geographic ontology with the specification of different topological relationships, generated with Protégé, in order to illustrate how mandatory semantic constraints are represented.

In the example in Figure 5, gas stations and bus stops must have a mandatory constraint with a road because every gas station and every bus stop must topologically ***touch*** one or more instances of a road. Roads, however, do not necessarily have gas stations or bus stops, so their relationship is not represented. Cities must also ***contain*** at least one road, while roads have no mandatory relationship with city. Notice in the OWL representation that minimum cardinality 1 is explicitly represented and can be easily retrieved.

To evaluate the amount of well-known dependences in real geo-ontologies we analyzed the first geo-ontology of Portugal, named geo-net-pt01 (Chaves et al., 2005b). Although not all elements of the geographic domain have been defined in geo-net-pt01, there are many ***one-one*** and ***one-many*** dependences.

The repository of the geo-ontology stores three levels of information: geo-administrative, geo-physical, and network. The geo-administrative level stores administrative information about territorial division, and includes geographic feature types such as municipalities, streets, and so forth. The network level stores nonspatial data and relationships about the geo-administrative layer (e.g., population of a district). The geo-physical level stores feature types including continents, oceans, lakes, bays, water bodies, and so forth.

In geo-net-pt01, among 58 different spatial feature types, 55 ***one-one*** relationships were defined in the geo-administrative level.

The following section introduces the problem of mining geographic data with well-known dependences.

THE PROBLEM OF GEOGRAPHIC DEPENDENCES IN SPATIAL ASSOCIATION RULE MINING

In transactional data mining, every row in the dataset to be mined is usually a transaction and columns are items, while in spatial data mining, every row is an instance (e.g., Buenos Aires) of a reference object type (e.g., city), called ***target feature type***, and columns are predicates. Every predicate is related to a nonspatial attribute (e.g., population) of the target feature type or a spatial predicate. Spatial predicate is a ***relevant feature type*** that is spatially related to specific instances of the target feature type (e.g., contains factory). Spatial predicates are extracted with operations provided by GIS, and can be represented at different granularity levels (Han & Fu, 1995; Lu, et al. 1993), according to the objective of the discovery. For example, chemical factory, metallurgical factory, and textile factory could be used instead of factory.

Spatial predicates are computed with spatial joins between all instances ***t*** of a target feature type ***T*** (e.g., city) and all instances ***o*** (e.g., Rio de la Plata) of every relevant feature type ***O*** (e.g.,

river) in a set of relevant feature types S (e.g., river, port, street, factory) that have any spatial relationship (e.g., touches, contains, close, far) with T. Being T a set of instances $T=\{t_1, t_2,\ldots,t_n\}$, $S = \{ O_1, O_2,\ldots, O_m\}$, and $O_i = \{ o_1, o_2,\ldots, o_q\}$, the extraction of spatial predicates implies the comparison of every instance of T with every instance of O, for all $O \subset S$.

The spatial predicate computation is the first step for extracting association rules from geographic databases. An association rule consists of an implication of the form $X \rightarrow Y$, where X and Y are sets of items co-occurring in a given tuple (Agrawal, Imielinkski & Swami, 1993). ***Spatial*** association rules are defined in terms of spatial predicates, where at least one element in X or Y is a spatial predicate (Koperski, 1995). For example, $is\_a(Slum) \wedge far\_from(WaterNetwork) \rightarrow disease=Hepatitis$ is a spatial association rule.

We assume that $F = \{f_1, f_2, \ldots, f_k, \ldots, f_n\}$ is a set of nonspatial attributes (e.g., population) and spatial predicates (e.g., close_to(Water)) that characterize a reference feature type, and Ψ (dataset) is a set of instances of a reference feature type**,** where each instance is a row W such that $W \subseteq F$. There is exactly one tuple in the dataset to be mined for each instance of the reference feature type.

The support s of a predicate set X is the percentage of tuples in which the predicate set X occurs as a subset. The support of the rule $X \rightarrow Y$ is given as $s(X \cup Y)$.

The rule $X \rightarrow Y$ is valid in Ψ with confidence factor $0 \leq c \leq$ **1, if at least** c% of the instances in Ψ that satisfy X also satisfy Y. The notation $X \rightarrow Y(c)$ specifies that the rule $X \rightarrow Y$ has confidence factor of c. More precisely, the confidence factor is given as $s(X \cup Y)/s(X)$.

The general problem of mining ***spatial*** association rules can be decomposed in three main steps, where the first one is usually performed as a data preprocessing method:

a. **Extract spatial predicates:** A spatial predicate is a spatial relationship (e.g., distance, order, topological) between the reference feature type and a set of relevant feature types.
b. **Find all frequent patterns/predicates:** A set of predicates is a frequent pattern if its support is at least equal to a certain threshold, called minsup.
c. **Generate strong rules:** A rule is strong if it reaches minimum support and the confidence is at least equal to a certain threshold, called minconf.

Assertion 1 (Agrawal & Srikant, 1994): if a predicate set Z is a frequent pattern, then every subset of Z will also be frequent. If the set Z is infrequent, then every set that contains Z is infrequent too. All rules derived from Z satisfy the support constraint if Z satisfies the support constraints.

Well-known geographic dependences appear in the three steps of the spatial association rule mining process. In the first step (a) well-known geographic dependences may exist among T and any $O \subset S$**.** In the second (b) and third (c) steps, dependences exist among relevant feature types, that is, between pairs of $O \subset S$. In the following sections we describe the problem that such dependences generate in frequent geographic pattern mining and what has been done so far to reduce this problem.

Geographic Dependences Between the Target Feature Type and Relevant Feature Types

In data preprocessing, time and effort are required from the data mining user to extract spatial relationships and transform geographic data in a single table or single file, which is the input format required by most data mining algorithms. Even in multirelational data mining where geographic data are transformed to first-order logic, the process of extracting spatial relationships is required.

The problem of ***which*** spatial relationships should be considered for knowledge discovery has been addressed in earlier works. (Koperski & Han, 1995;Lu et al., 1993) presented a top-down progressive refinement method where spatial approximations are calculated in a first step, and in a second step, more precise spatial relationships are computed to the outcome of the first step. The method has been implemented in the module Geo-Associator of the GeoMiner system (Han, Koperski & Stefanvic, 1997), which is no longer available. Ester et al., (2000) proposed new operations such as graphs and paths to compute spatial neighborhoods. However, these operations are not implemented by most GIS, and to compute all relationships between all objects in the database in order to obtain the graphs and paths is computationally expensive for real databases. Appice et al., (2005) proposed an upgrade of Geo-Associator to first-order logic, and all spatial relationships are extracted. This process is computationally expensive and nontrivial in real databases. While the above approaches consider different spatial relationships and any geometric object type, a few approaches such as (Huang, Shekhar & Xiong, 2004; Yoo & Shekhar, 2006) compute only distance relationships for point object types.

Table 1 shows an example of a spatial dataset at a high granularity level, where every row is a city and predicates refer to different geographic object types (port, water body, hospital, street, and factory) spatially related to city. Let us consider two geographic dependences: city and street, and port and water body, where the former is between the target feature type and a relevant feature type and the latter is among the two relevant feature types.

In the dataset shown in Table 1, the dependence between the target feature type city and the relevant feature type street is explicit, because every city has at least one street and the predicate ***contains***(***Street***) has a 100% support. Predicates with 100% support appear in at least half of the total number of patterns and generate a large number of noninteresting association rules. For example, a rule such as ***contains(factory)*** → ***contains(Street)*** expresses that cities that contain factories do also contain streets. Although such a rule seems to be interesting, it can be considered obvious due the simple fact that ***all*** cities contain streets, having they factories, or not.

Table 1. Example of a preprocessed dataset in a high granularity level for mining frequent patterns and SAR

Tuple (city)	Spatial Predicates
1	contains(Port), contains(Hospital), contains(Street), contains(Factory), crosses(Water Body)
2	contains(Hospital), contains(Street), crosses(Water Body)
3	contains(Port), contains(Street), contains(Factory), crosses(Water Body)
4	contains(Port), contains(Hospital), contains(Street), crosses(Water Body)
5	contains(Port), contains(Hospital), contains(Street), contains(Factory), crosses(Water Body)
6	contains(Hospital), contains(Street), contains(Factory)

Table 2. Frequent patterns and rules with dependences

Min Sup %	Total FrequentSets/ Rules	Rules with Dependence / Rules without Dependence	FrequentSets with dependence / FrequentSets without dependence
20	31 / 180	130 / 50	16 / 15
50	25 / 96	72 / 24	13 / 12

Table 2 shows the result of a small experiment performed with Apriori (Agrawal & Srikant, 1994) over the dataset in Table 1. Considering 20% minimum support, 31 frequent sets and 180 rules were generated. Among the 31 frequent sets and the 180 rules, 16 frequent sets and 130 rules had the dependence ***contains(Street)***. Notice that increasing minimum support to 50% does not warrant the elimination of the geographic dependence. Although the number of frequent sets is reduced to 25 and rules to 96, 13 frequent sets and 72 rules still have the dependence.

Geographic dependences besides generating a large number of well-known patterns and association rules, require unnecessary spatial joins. To illustrate the power that semantics may have in spatial join computation, let us consider a few examples, shown in Table 3. Without considering semantics, ***all*** topological relationships between two spatial feature types would be tested in order to verify which one holds. Considering semantics, the number of relationships to test reduces significantly. As shown in Table 3, the only topological relationship semantically consistent between gas station and road should be ***touches.*** A city hall must be ***within*** a city, while a water body can be ***disjoint***, ***touch***, or ***cross*** a road.

Although the topological relationships shown in Table 3 are semantically possible, not all of them are interesting for knowledge discovery. So, if besides considering the semantics of spatial features we also consider spatial integrity constraints, it is possible to reduce still further the number of topological relationships and define which should be computed for knowledge discovery. Remembering that ***mandatory*** relationships produce well known patterns and that only ***possible*** relationships are interesting for knowledge discovery, Table 4 shows the topological relationships of the same objects in Table 3 that would be computed if semantics and integrity constraints were considered. The pairs gas station and road, bridge and water body, city hall and city, as well as treated water net and city have ***mandatory*** one-one or one-many constraints and no relationship is necessary for KDD.

Despite ***mandatory*** and ***prohibited*** constraints do not explicitly define the interesting spatial

Table 3. Possible and mandatory topological relationships considering semantics of feature types

Topological Relationship / Semantic Combinations	Disjoint	Overlaps	Touches	Contains	Within	Crosses	Equals
Gas Station and Road			✓				
Bridge and Water Body						✓	
City Hall and City					✓		
Water Body and Road	✓		✓			✓	
Treated Water Net and City			✓		✓	✓	

Table 4. Possible topological relationships for knowledge discovery

Topological Relationship / Semantic Combinations	Disjoint	Overlaps	Touches	Contains	Within	Crosses	Equals
Gas Station and Road							
Bridge and Water Body							
City Hall and City							
Water Body and Road	✓		✓			✓	
Treated Water Net and City							

relationships to be extracted for knowledge discovery, we are able to eliminate those which are either ***mandatory*** or ***prohibited***, and specify those which are ***possible,*** as will be explained in Section 4.

Geographic Dependences Among Relevant Feature Types

To find ***frequent predicate sets*** and ***extract strong association rules,*** predicates are combined with each other for the different instances of the target feature type ***T,*** and not among ***T*** and ***O*** as explained in the previous section.

To illustrate the geographic dependence replication process in frequent geographic pattern mining, let us consider the frequent set generation introduced by (Agrawal & Srikant, 1994) for the Apriori algorithm. Apriori performs multiple passes over the dataset. In the first pass, the support of the individual elements is computed to determine ***k***-predicate sets. In the subsequent passes, given ***k*** as the number of the current pass, the large sets L_{k-1} in the previous pass (***k*** -1) are grouped into sets C_k with ***k*** elements, which are called ***candidate sets***. The support of each candidate set is computed, and if it is equal or higher than minimum support, then this set is considered frequent/large. This process continues until the number of large sets is zero.

Geographic dependences appear the first time in frequent sets with 2 elements, where ***k***=2. Table 5 shows the frequent sets extracted from the dataset in Table 1 with 50% minimum support, where ***k*** is the number of elements in the frequent sets. Notice that since the dependence has minimum support, that is, a frequent predicate set, this dependence is replicated to many frequent sets of size ***k>2*** with predicates that reach minimum support, as shown in bold style in Table 5. Considering such a small example and high minimum support, one single geographic dependence participates in six frequent sets, which represents 30% of the frequent sets. Notice that the number of rules having a geographic dependence will be much larger than the frequent sets, mainly when the largest frequent set (with 4 elements) contains the dependence.

In Table 5, we can observe that the technique of generating ***closed frequent sets*** (Paskier et al., 1999; Zaki & Hsiao, 2002) would not eliminate geographic dependences, because both sets with 4 elements that contain the dependence are closed frequent sets. The closed frequent set approach

Table 5. Large predicate sets with 50% minimum support

k	*Frequent sets with support 50%*
1	{contains(Port)}, {contains(Hospital)}, {contains(Street)}, {contains(Factory)}, {crosses(WaterBody)}
2	{contains(Port),contains(Hospital)}, {contains(Port),contains(Street)}, {contains(Port),contains(Factory)}, ***{contains(Port),crosses(WaterBody)},*** {contains(Hospital),contains(Street)}, {contains(Hospital),contains(Factory)}, {contains(Hospital),crosses(WaterBody)}, {contains(Street),contains(Factory)}, {contains(Street),crosses(WaterBody)}, {contains(Factory),crosses(WaterBody)}
3	{contains(Port),contains(Hospital),contains(Street)}, ***{contains(Port),contains(Hospital),crosses(WaterBody)},*** ***{contains(Port),contains(Street),crosses(WaterBody)},*** ***{contains(Port),contains(Factory),crosses(WaterBody)},*** {contains(Port),contains(Street),contains(Factory)}, {contains(Hospital),contains(Street),contains(Factory)} {contains(Hospital),contains(Street),crosses(WaterBody)}, {contains(Street),contains(Factory),crosses(WaterBody)}
4	***{contains(Port),contains(Hospital),contains(Street),crosses(WaterBody)}*** ***{contains(Port),contains(Street),contains(Factory),crosses(WaterBody)}***

eliminates ***redundant frequent sets***, but does not eliminate well known dependences if applied to the geographic domain.

In order to evaluate the amount of well-known rules generated with the dependence, let us observe Table 6, which shows a few examples of association rules generated with frequent predicate sets of size 2 {***Contains***(***Port***),***crosses***(***Water Body***)}, size 3 {***Contains***(***Port***),***contains***(***Hospital***),***crosses***(***Water Body***)}, and size 4 {***Contains***(***Port***), ***contains***(***Hospital***), ***contains***(***Street***),***crosses***(***Water Body***)}. Rules 1 and 2 are generated from the set with two elements, and represent a single geographic dependence and its inverse. Rules 3, 4, 5, and 6 reproduce rules 1 and 2 with an additional element in the antecedent or the consequent of rule. The same happens with frequent sets that contain 4 elements. Rules 7, 8, and 9 are rules 1 and 2 with two additional elements that combined with the dependence reached minimum support.

Approaches that reduce the number of rules and eliminate redundant rules (Zaki, 2000) do not warrant the elimination of all association rules that contain geographic dependences.

Existing algorithms for mining frequent geographic patterns and generating strong spatial association rules do neither make use of semantic knowledge to specify which spatial relationships should be computed in data preprocessing, nor to reduce the number of well-known patterns. Koperski and Han (1995) reduces the number of rules using minimum support during the predicate generation. Clementini et al., (2000) presented a similar method for mining association rules from geographic objects with broad boundaries. Appice et al., (2005) reduces the number of rules with user specified ***pattern constraints***, which require a lot of background knowledge from the data mining user. This method is inefficient since pattern constraints are applied in post-processing steps, after both frequent sets and association rules have already been generated.

Because of the dependence replication process in both frequent sets and association rules, shown in Table 5 and Table 6 respectively, it might be difficult for the data mining user to analyze all rules to discover if they are really interesting or not. To help the data mining user, in the following section we present a framework to remove all well known geographic dependences, warranting that no association rules with such dependences will be generated.

A FRAMEWORK FOR GEOGRAPHIC DATA PREPROCESSING AND

Table 6. Examples of association rules with frequent sets of size 2, 3, and 4 having the geographic dependence

Set Size	*Rule*	*Possible Rules*
k=2	1	contains(Port) → crosses(Water Body)
k=2	2	crosses(Water Body) → contains(Port)
k=3	3	contains(Hospital) ^ contains(Port) → crosses(Water Body)
k=3	4	contains(Hospital) ^ crosses(Water Body) → contains(Port)
k=3	5	contains(Hospital) → contains(Port) ^ crosses(Water Body)
k=3	6	contains(Port) ^ crosses(Water Body) → contains(Hospital)
k=4	7	contains (Street) ^ ***contains(Port)*** → ***crosses(Water Body)*** ^ contains (Hospital)
k=4	8	contains (Street) → ***contains(Port)*** ^ ***crosses(Water Body)*** ^ intersects (Hospital)
k=4	9	contains (Street) ^ intersects (Hospital) → ***contains(Port)*** ^ ***crosses (Water Body)***

SPATIAL ASSOCIATION RULE MINING WITH ONTOLOGIES

Recently, in (Bogorny et al., 2005b, 2006a, 2006b, 2006c) we introduced the idea of using semantic knowledge for reducing spatial joins and well known patterns in SAR mining. In Bogorny et al., (2006a) we proposed to eliminate well-known patterns among the target feature type and relevant feature types with intelligent geographic data preprocessing. In data preprocessing, however, not all well-known dependences can be removed. Then, we presented a frequent pattern mining algorithm that uses semantic knowledge to eliminate dependences among relevant feature types during the frequent set generation (Bogorny et al., 2006b). In Bogorny et al., (2006c) we proposed an integrated framework, which eliminates geographic dependences completely in both data preprocessing and frequent pattern generation, using geographic database schemas as prior knowledge.

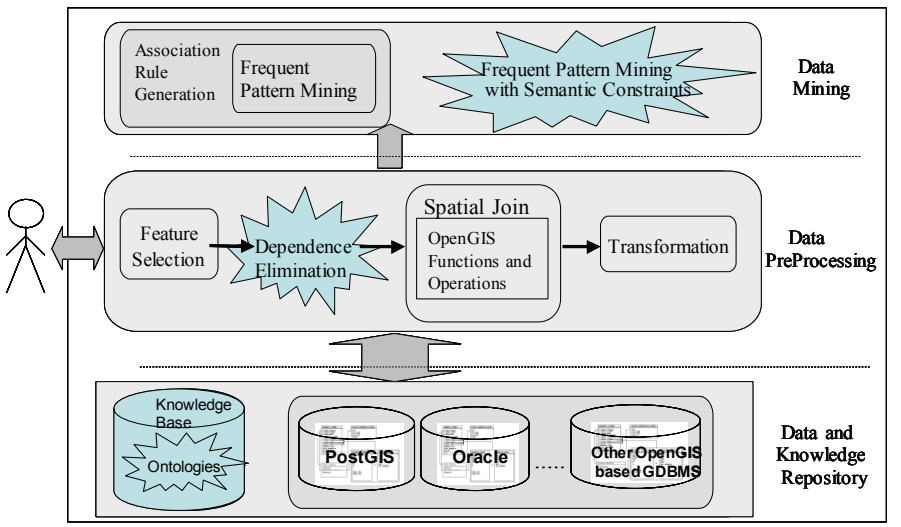
This section presents an interoperable framework for geographic data preprocessing and spatial association rule mining using geographic ontologies. Ontologies are used not only to eliminate well known dependences, but to verify which spatial relationships should be computed in the spatial predicate computation.

Figure 6 shows the framework that can be viewed in three levels: data repository, data preprocessing, and data mining. At the bottom are the geographic data repositories: the knowledge repository which stores geo-ontologies and geographic databases stored in GDBMS (geographic database management systems) constructed under OGC specifications. Following the OGC specifications (Open GIS Consortium, 1999b) makes our framework interoperable with all GDBMS constructed under OGC specifications (e.g., Oracle, PostGIS, MySQL, etc).

At the center is the spatial data preparation level, which covers the ***gap*** between data mining tools and geographic databases. At this level, data and knowledge repositories are accessed through JDBC/ODBC connections and data are retrieved, preprocessed, and transformed into the single table format. At this level, dependences among the target feature and relevant features are removed, as described in the next section.

On the top are the data mining toolkits or algorithms for mining frequent patterns and

Figure 6. A Framework for mining frequent geographic patterns using ontologies



generating association rules. At this level, a new method for mining frequent geographic patterns is presented. Dependences among relevant feature types that can only be removed into the data mining algorithm are eliminated during the frequent set generation, as will be explained along with this section.

Data Preprocessing: Using Semantics to Eliminate Geographic Dependences between the Target Feature Type and the Relevant Feature Types

There are four main steps to implement the tasks of geographic data preprocessing for association rule mining: ***Feature Selection, Dependence Elimination, Spatial Join, a***nd ***Transformation.*** The ***Feature Selection*** step retrieves all relevant information from the database such that the user can choose the target feature type ***T,*** the target feature nonspatial attributes and the set ***S*** of relevant feature types that may have some influence on ***T***. The feature types as well as their geometric attributes are retrieved through the OpenGIS database schema metadata, stored in the relation ***geometry_columns*** (see Bogorny et al., 2005a) for details.

The algorithm that implements the remaining data preprocessing steps is presented in Figure 7. The ***Dependence Elimination*** step searches the ontology φ and verifies the properties of ***T.*** If ***T*** has a mandatory dependence ***M*** with any ***O*** in ***S***, then ***O*** 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. If a prohibited relationship ***P*** is defined between ***T*** and ***O*** in the ontology φ, then the set of possible relationships to compute for data mining is given by $\boldsymbol{D}_{(T,O)} = \boldsymbol{R} - \boldsymbol{P}_{(T,O)}$, where ***R*** is the set of all topological relationships ***R = {touches, contains,***

Figure 7. Pseudo-code of the data preprocessing algorithm

```
Given:
    GDB, // geographic database
    φ, // geographic ontology
    T, // target feature type
    S, // set of relevant feature types O
    R; // set of all topological relationships
Variables:
    D; // relationships to compute for Data mining

Find: a dataset Ψ without geographic dependences between T and S;

Method:
Dependence_Elimination
Begin
    Ψ = T – geometry column;
    For (i=1; i=#O in S, i++) do
    Begin
        Find T in φ;
        If (T has a one-one or one-many property with Oi in φ)
            Remove Oi from S;   // dependence elimination
        Else
            If (T has prohibited properties P with Oi in φ)
                D = R – P; // possible relationships to compute
            Else
                D = R // all topological relationships
        Ψ = Ψ + Spatial_Join (D,T,Oi);//computes spatial relationships D between T and O
    End;
End;
Transformation (Ψ) // transforms the resultant dataset into the data mining algorithm
                   // format preserving the non-spatial attributes of T;
```

within, crosses, overlaps, equals, disjoint}. If there is no property of ***T*** in φ that relates ***T*** and ***O***, then all relationships are computed.

The ***Spatial Join*** step computes the spatial relationships ***D*** between ***T*** and all remaining ***O*** in ***S.*** Spatial joins ***D*** to extract spatial predicates are performed on-the-fly with operations provided by the GIS.

The ***Transformation*** step transposes as well as discretizes the ***Spatial Join*** module output (Ψ) into the single table format understandable to association rule mining algorithms.

Frequent Pattern Generation: Using Semantics to Eliminate Geographic Dependences Among Relevant Features

Frequent pattern and association rule mining algorithms, under rare exceptions (Han, Pei & Yin, 2000) generate candidates and frequent sets. The candidate generation in spatial data mining is not a problem because the number of predicates is much smaller than the number of items in transactional databases (Shekhar & Chawla, 2003). Moreover, the computational cost relies on the spatial join computation.

Approaches that generate closed frequent sets do previously compute the frequent sets, and than verify if they are closed. Although they reduce the number of frequent sets, they do not warrant the elimination of well known geographic patterns. In SAR mining, it is more important to reduce the number of frequent sets than warrant that the resultant frequent sets are free of well-known dependences, aiming to generate more interesting frequent sets.

Apriori (Agrawal & Srikant, 1994) has been the basis for dozens of algorithms for mining spatial and nonspatial frequent sets, and association rules. We will illustrate the method of geographic dependence elimination during the frequent set generation using Apriori, as shown in Figure 8.

We propose to remove from the candidate sets all pairs of elements that have geographic dependences. As in Apriori, multiple passes are performed over the dataset. In the first pass, the support of the individual elements is computed to determine large-predicate sets. In the subsequent passes, given ***k*** as the number of the current pass, the large/frequent sets $\boldsymbol{L}_{k-1}$ in the previous pass (***k***-1) are grouped into sets $\boldsymbol{C}_k$ with ***k*** elements, which are called ***candidate sets***. Then the support of each candidate set is computed, and if it is equal or higher than minimum support, then this set is considered frequent. This process continues until the number of frequent sets is zero.

Similarly to Srikant and Agrawal (1995),

Figure 8. Frequent set generation function

```
Given: ϕ, Ψ, minsup;
L1 = {large 1-predicate sets};
For ( k = 2; Lk-1 != ∅; k++ ) do begin
  Ck = apriori_gen(Lk-1); // Generates new candidates
  If (k=2)
      Forall candidates c ∈ C2 do
          If (feature types in pair c have any one-one or one-many relationship in ϕ)
              Remove the subset c from C2.
  Forall rows w ∈ Ψ do begin
    Cw = subset(Ck, w); // Candidates contained in w
    Forall candidates c ∈ Cw do
      c.count++;
  End;
  Lk = {c ∈ Ck | c.count ≥ minsup};
End;
Answer = ∪k Lk
```

which eliminates in the second pass candidate sets that contain both parent and child specified in concept hierarchies, we eliminate all candidate sets which contain geographic dependences, but independently of any concept hierarchy.

The dependences are eliminated in an efficient way, when generating candidates with 2 elements, and before checking their frequency. If the pairs of predicates (e.g., ***contains(Port), contains(Water Body)***) contain feature types (e.g., ***Port, Water Body***) that have a mandatory constraint in the ontology φ**,** then all pairs of predicates with a dependence in φ are removed from $\boldsymbol{C}_2$.

According to Assertion 1, this step ***warrants*** that the pairs of geographic objects that have a mandatory constraint in the ontology φ will neither appear together in the frequent sets, nor in the spatial association rules. This makes the method effective independently of other thresholds, and clearly improves in efficiency, since less frequent sets will be generated.

The main strength of this method in our framework is its simplicity. This single, but very effective and efficient step, removes all well-known geographic dependences, and can be implemented by any algorithm that generates frequent sets. Considering the example of frequent sets shown in Table 5, the dependence is eliminated when it appears at the first time, such that no larger frequent sets or association rules with the dependence will be generated.

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. Districts, a database table with 109 polygons and nonspatial attributes, such as population and sanitary condition, was defined as the target feature type ***T.*** Datasets with different relevant feature types (e.g., bus routes—4062 multilines, slums —513 polygons, water resources—1030 multilines, gas stations 450 points) were preprocessed and mined, using ontologies and without using ontologies.

Estimating the time reduction to compute spatial joins for mining frequent patterns is very difficult, since this step is completely data dependent. The computational time reduction to extract spatial joins depends on three main aspects: how many dependences (relevant feature types) are eliminated in data preprocessing; the geometry type of the relevant feature (point, line, or polygon); and the number of instances of the eliminated feature type (e.g., 60,000 rows). For example, if a relevant feature type with 57 580 polygons is eliminated, spatial join computation would significantly decrease. If the eliminated feature type has 3062 points, time reduction would be less significant. However, for every relevant feature type eliminated, no spatial join is necessary, and this warrants preprocessing time reduction.

To evaluate the frequent pattern reduction by pruning the input space, Figure 9 describes an experiment performed with Apriori, where 2 dependences between the reference object type and the relevant feature types were eliminated. Notice that input space pruning reduces frequent patterns independently of minimum support. Considering ***minsup*** 10%, 15%, and 20%, the elimination of one single dependence pruned the frequent sets around 50%. The elimination of two dependences reduced the number of frequent sets in 75%. The rule reduction is still more significant, as can be observed in Figure 10, reaching around 70% when one dependence is removed and 90% when two dependences are eliminated, independently of minimum support.

Algorithms that generate closed frequent sets and eliminate nonredundant rules can reduce still further the number of both frequent sets and association rules if applied to the geographic domain using our method for pruning the input space.

Figure 11 shows the result of an experiment where two dependences among relevant feature types were eliminated during the frequent set

Figure 9. Frequent sets generated with input space pruning

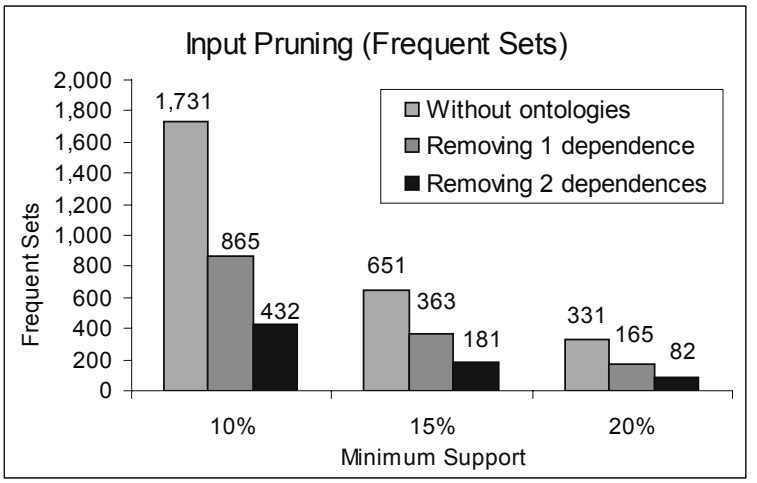
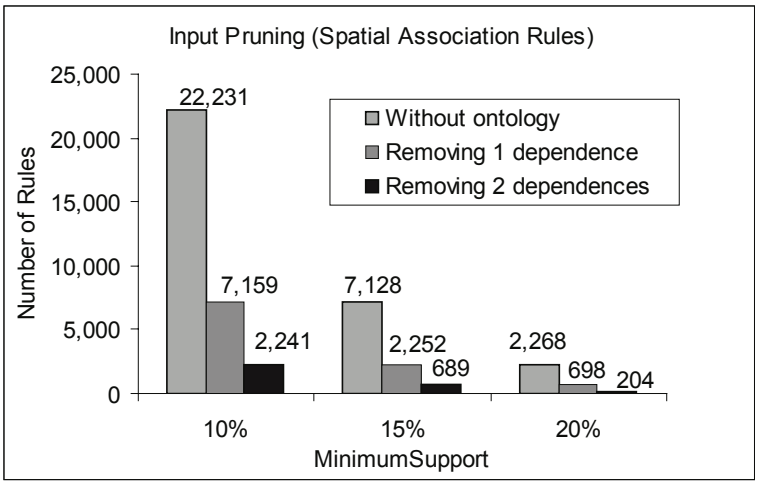


Figure 10. Spatial association rules with input space pruning and 70% minimum confidence



generation, but without input pruning. Notice that even using ontologies only in the frequent set generation we get a reduction on the number of frequent sets independently of minimum support. Moreover, the higher the number of dependences, the more significant is the reduction.

Figure 12 shows an experiment where dependences were eliminated in both input space (between the target feature and relevant features) and during the frequent set generation (among relevant features). The total number of frequent sets is reduced in more than 50% by removing one single dependence, independently of minimum support. Using ontologies we completely eliminate well known dependences, and very efficiently.

FUTURE TRENDS

Data mining techniques to extract knowledge

Figure 11. Frequent sets generated with frequent set pruning

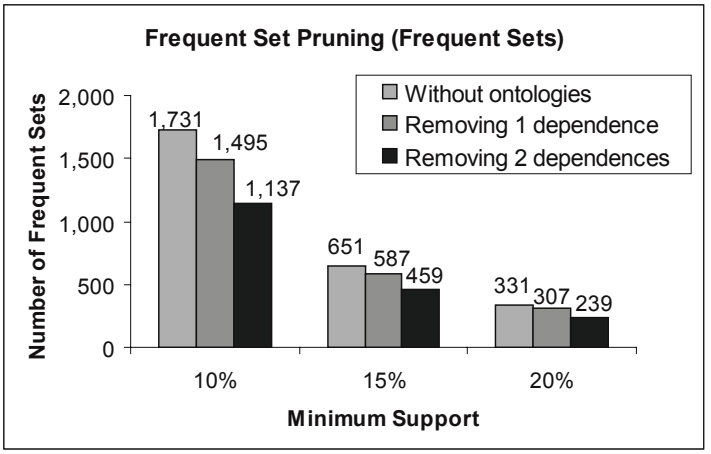
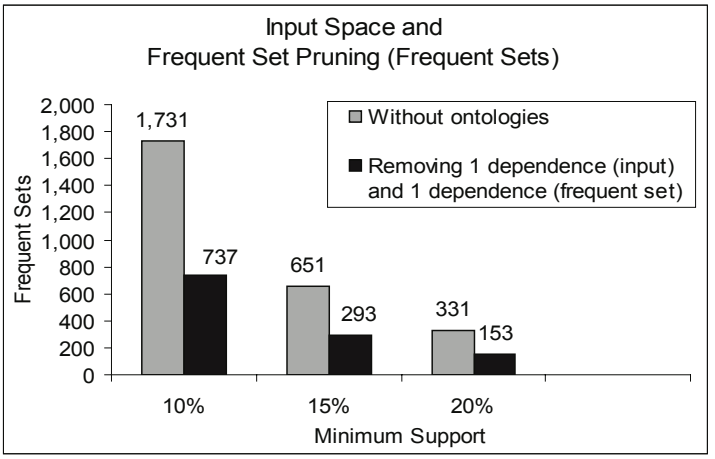


Figure 12. Frequent sets generated with input space and frequent set pruning



from large spatial and nonspatial databases have mainly considered syntactic constraints and the data by itself, without considering semantics. The result is that the same geographic dependences that are well-known by GDB designers and explicitly represented in GDB schemas and geo-ontologies to warrant the consistency of the data, are extracted by data mining algorithms, which should discover only novel and useful patterns. When dealing with geographic data, which are semantically interdependent because of their nature, the meaning of data needs to be considered, at least to avoid the extraction of well known patterns.

There is an emerging necessity to consider semantic geographic domain knowledge in spatial data mining. The large amount of knowledge explicitly represented in geographic database schemas and spatio-temporal ontologies needs to be incorporated into data mining techniques, since they provide a valuable source of domain knowledge. How to use this knowledge in data mining systems and for which purposes are still open problems. In this chapter, we presented an

efficient solution, addressing a small fraction of these problems. We used geo-ontologies in spatial association rule mining to reduce well-known patterns, but the use of ontologies in different data mining techniques such as clustering, classification, and outlier detection are still open problems. In clustering, for example, the use of semantics could either avoid the separation of geographic objects that have mandatory constraints or organize them into the same cluster without the need of computing their relationship. The use of prior knowledge to evaluate the interestingness of patterns extracted with the different data mining techniques still needs to be addressed.

The development of toolkits that integrate data mining techniques, geographic databases, and knowledge repositories is another need for practical applications. Although a large number of algorithms has been proposed, their implementation in toolkits with friendly graphical user interfaces that cover the whole KDD process is rare. The gap between data mining techniques and geographic databases is still a problem that makes geographic data preprocessing be the most effort and time consuming step for knowledge discovery in these databases.

CONCLUSION

This chapter presented an intelligent framework for geographic data preprocessing and SAR mining using geo-ontologies as prior knowledge. The knowledge refers to mandatory and prohibited semantic geographic constraints, which are explicitly represented in geo-ontologies because they are part of the concepts of geographic data. We showed that explicit mandatory relationships produce irrelevant patterns, and that prohibited relationships do not need to be computed, since they will never hold if the database is consistent. Possible implicit spatial relationships may lead to more interesting patterns and rules, and they can be inferred using geo-ontologies.

Experiments showed that independent of the number of elements, one dependence is enough to prune a large number of patterns and rules, and the higher the number of eliminated semantic constraints, the larger is the frequent pattern and rule reduction. We showed that well-known dependences can be partially eliminated with intelligent data preprocessing, independently of the algorithm to be used for frequent pattern mining. To completely eliminate geographic dependences we presented a pruning method that can be applied to any algorithm that generates frequent sets, including closed frequent sets. Algorithms for mining nonredundant association rules can reduce the number of rules further if applied to the geographic domain using our method to generate frequent sets.

Considering semantics in geographic data preprocessing and frequent pattern mining has three main advantages: spatial relationships between feature types with dependences are not computed; the number of both frequent sets and association rules is significantly reduced; and the most important, the generated frequent sets and rules are free of associations that are previously known as noninteresting.

The main contribution of the method presented in this chapter for mining spatial association rules is for the data mining user, which will analyze much less obvious rules. The method is effective independently of other thresholds, and warrants that geographic domain associations will not appear among the resultant set of rules.

ACKNOWLEDGMENT

The authors would like to thank both CAPES and CNPq, which partially provided the financial support for this research. To Procempa, for the geographic database and to Nodo XLDB da Linguateca of Universidade de Lisboa for the geographic ontology. Our special thanks for Mariusa Warpechowski and Daniela Leal Musa

for the ontology modeling support, and for Sandro da Silva Camargo for the support with data mining algorithms.

REFERENCES

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Buneman, P. & Jajodia, S. (Eds.), *ACM SIGMOD International Conference on Management of Data: 20,* 207-216. New York: ACM Press.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In Bocca, J. B., Jarke, M., & Zaniolo, C. (Eds.), *International Conference on Very Large Databases, 20,* 487-499. San Francisco: Morgan Kaufmann Publishers Inc.

Appice, A., Berardi, M., Ceci, M., & Malerba, D. (2005). Mining and filtering multilevel spatial association rules with ARES. In M. Hacid, N. V. Murray, Z. W. Ras, S. Tsumoto (Eds.), *Foundations of Intelligent Systems, 15th International Symposium ISMIS. Vol. 3488*. (pp. 342-353). Berlin: Springer.

Bogorny, V. & Iochpe, C. (2001). Extending the opengis model to support topological integrity constraints. In Mattoso, M. & Xexéo, G. (Eds.), *16th Brazilian Symposium in Databases* (pp. 25-39). Rio de Janeiro: COPPE/UFRJ.

Bogorny, V., Engel, P. M., & Alvares, L.O. (2005a). A reuse-based spatial data preparation framework for data mining. In J. Debenham, K. Zhang (Eds.), *15th International Conference on Software Engineering and Knowledge Engineering* (pp. 649-652). Taipei: Knowledge Systems Institute.

Bogorny, V., Engel, P. M., & Alvares, L.O. (2005b). Towards the reduction of spatial join for knowledge discovery in geographic databases using geo-ontologies and spatial integrity constraints. In Ackermann, M., Berendt, B., Grobelink, M., & Avatek, V. (Eds.), *ECML/PKDD 2nd Workshop on Knowledge Discovery and Ontologies,* (pp. 51-58). Porto.

Bogorny, V., Engel, P. M., & Alvares, L.O. (2006a). GeoARM: An interoperable framework to improve geographic data preprocessing and spatial association rule mining. *In 18th International Conference on Software Engineering and Knowledge Engineering* (pp. 70-84). San Francisco: Knowledge Systems Institute.

Bogorny, V., Camargo, S., Engel, P., M., & Alvares, L.O. (2006b). Towards elimination of well known geographic domain patterns in spatial association rule mining. *In 3rd IEEE International Conference on Intelligent Systems* (pp. 532-537). London: IEEE Computer Society.

Bogorny, V., Camargo, S., Engel, P., & Alvares, L. O. (2006c). Mining frequent geographic patterns with knowledge constraints. *In 14th ACM International Symposium on Advances in Geographic Information Systems.* Arlington, November (to appear).

Chaves, M. S., Silva, M. J., & Martins, B. (2005a). A geographic knowledge base for semantic web applications. In Heuser, C. A. (Ed.), *20th Brazilian Symposium on Databases* (pp. 40-54). Uberlandia: UFU.

Chaves, M. S., Silva, M. J., & Martins, B. (2005b). *GKB—Geographic Knowledge Base*. (TR05-12). DI/FCUL.

Clementini, E., Di Felice, P., & Van Ostern, P. (1993). A small set of formal topological relationships for end-user interaction. In D.J. Abel, B.C. Ooi (Eds.), *Advances in Spatial Databases, 3rd International Symposium, 692*, 277-295. Singapure: Springer.

Cockcroft, S. (1997). A Taxonomy of spatial data integrity constraints. *Geoinformatica, 1*(4), 327-343.

Clementini, E., Felice, Di, P., & Koperski, K. (2000). Mining multiple-level spatial association rules for objects with a broad boundary. *Data & Knowledge Engineering, 34*(3), 251–270.

Egenhofer, M. & Franzosa, R. (1995). On the equivalence of topological relations. *International Journal of Geographical Information Systems, 9*(2), 133-152.

Ester, M., Frommelt, A., Kriegel, H.-P., & Sander, J. (2000). Spatial data mining: database primitives, algorithms and efficient DBMS support. *Journal of Data Mining and Knowledge Discovery, 4*(2-3), 193-216.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to discovery knowledge in databases. *AI Magazine, 3*(17), 37-54.

Gruber, T. R. (1993). Towards principles for the design of ontologies used for knowledge sharing. Formal ontology in conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies, 43,* 907-928.

Guarino, N. (1998). Formal ontology and information systems. In N. Guarino (Ed.), *International Conference on Formal Ontology in Information Systems* (pp. 3-15). Trento: IOS Press.

Han, J. & Fu, Y. (1995). Discovery of multiple-level association rules from large databases. In U. Dayal, P.M.D. Gray, S. Nishio (Eds), *International Conference on Very Large Data Bases* (pp. 420–431). Zurich: Morgan-Kaufmann.

Han, J., Koperski, K., & Stefanvic, N. (1997). GeoMiner: a system prototype for spatial data mining. In J. Peckham (Ed.), *ACMSIGMOD International Conference on Management of Data, 26,* 553-556. Tucson: ACM Press.

Han, J., Pei J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. In Chen, J.F. Naughton, P.A. Bernstein (Eds.)*, 20th ACMSIGMOD International Conference on Management of Data* (pp. 1-12) Dallas: ACM.

Huang, Y., Shekhar, S., & Xiong, H. (2004). Discovering co-location patterns from spatial datasets: A general approach. *IEEE Transactions on Knowledge and Data Engineering, 16*(12), 1472-1485.

Koperski, K., & Han, J. (1995). Discovery of spatial association rules in geographic information databases. In M.J. Egenhofer, J.R. Herring (Eds.), *4th International Symposium on Large Geographical Databases, 951,* 47-66. Portland: Springer.

Lu, W., Han, J., & Ooi, B. C. (1993). Discovery of general knowledge in large spatial databases. *In Far East Workshop on Geographic Information Systems*, (pp. 275-289). Singapure.

Open Gis Consortium. (1999a). T*opic 5, the OpenGIS abstract specification—OpenGIS features—Version 4*. Retrieved August 20, 2005, from http://www.OpenGIS.org/techno/specs.htm

Open Gis Consortium. (1999b). *Open GIS Simple Features Specification For SQL*. Retrieved August 20, 2005, from http://www.opengeospatial.org/specs

Open Gis Consortium. (2001). *Feature Geometry*. Retrieved August 20, 2005, from http://www.opengeospatial.org/specs.

Pasquier, N. Bastide, Y., Taouil, R., & Lakhal, L (1999). In Beeri, C., Buneman, P. (Eds.), *7th International Conference on Database Theory,1540,* 398-416. Jerusalem: Springer.

Servigne, S., Ubeda, T., Puricelli, A., & Laurini, R. (2000). A Methodology for spatial consistency improvement of geographic databases. *Geoinformatica, 4*(1), 7-34.

Shekhar, S. & Chawla, S. (2003). *Spatial databases: a tour.* Upper Saddle, NJ: Prentice Hall.

Srikant, R. & Agrawal, R. (1995). Mining generalized association rules. In U. Dayal, P. M. D.

Gray, S. Nishio (Eds.), *Proceedings of the 21st International Conference on Very Large Databases,* (pp.407-419). Zurich: Morgan Kaufmann.

Yoo, J. S., & Shekhar, S. (2006). A join-less approach for mining spatial co-location patterns. *IEEE Transactions on Knowledge and Data Engineering, 18*(10).

Zaki. M. (2000). Generating nonredundant association rules. In S.J. Simoff, O. R. Zaïane (Eds.), *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* (pp. 34-43) Boston: ACM Press.

Zaki., M., & Hsiao, C. (2002). CHARM: An efficient algorithm for closed itemset mining. In R.L. Grossman, J. Han, V. Kumar, H. Mannila, R. Motwani (Eds.), *Proceeding of the 2nd SIAM International Conference on Data Mining,* (pp. 457-473). Arlington: SIAM.

ADDITIONAL READING

Bernstein, A. Provost, Foster J.& Hill, S. (2005) Toward Intelligent Assistance for a Data Mining Process: An Ontology-Based Approach for Cost-Sensitive Classification. *IEEE Transactions on Knowledge and. Data Engineering. 17*(4), 503-518.

Bogorny, V., Valiati, J. F., Camargo, S. S., Engel, P. M., Kuijpers, B. & Alvares, L.O. (2006). Mining Maximal Generalized Frequent Geographic Patterns with Knowledge Constraints. *Sixth IEEE International Conference on Data Mining* (pp. 813-817). Hong Kong: IEEE Computer Society.

Bogorny, V. (2006). *Enhancing spatial association rule mining in geographic databases.* PhD Thesis. Porto Alegre, Brazil: Instituto de Informatica—UFRGS.

Chen, X., Zhou, X., Scherl, R.B. & Geller, J. (2003). Using an Interest Ontology for Improved Support in Rule Mining. In Y. Kambayashi, M. K. Mohania, W. Wolfram (Eds), *Fifth International Conference on Data WareHouse and Knowledge Discovery* (pp. 320-329). Prague: Springer.

Farzanyar, Z., Kangavari, M. & Hashemi, S. (2006). A New Algorithm for Mining Fuzzy Association Rules in the Large Databases Based on Ontology. *Workshops Proceedings of the 6th IEEE International Conference on Data Mining* (pp.65-69). Hong Kong: IEEE Computer Society.

Jozefowska, J., Lawrynowicz, A. & Lukaszewski, T. (2006). Frequent pattern discovery from OWL DLP knowledge bases. In S. Staab and V. Sv (Eds). *International Conference on Managing Knowledge in a World of Networks.* (pp. 287-302). Czech Republic: Springer.

Knowledge Discovery and Ontologies. (2004). ECML/PKDD Workshop. Retrieved February 12, 2006, from http://olp.dfki.de/pkdd04/cfp.htm.

Knowledge Discovery and Ontologies. (2005). ECML/PKDD Workshop. Retrieved February 12, 2006, from http://webhosting.vse.cz/svatek/KDO05

Mennis, J. Peuquet, D. J. (2003) The Role of Knowledge Representation in Geographic Knowledge Discovery: A Case Study. *Transactions in GIS, 7*(3), 371–391.

Singh, P. & Lee, Y. (2003) Context-Based Data Mining Using Ontologies. In I. Song, S. W. Liddle, T. Wang Ling, P. Scheuermann (Eds), *International Conference on Conceptual Modeling* (pp. 405-418). Chicago: Springer.

Xu, W. & Huang, H. (2006). Research and Application of Spatio-temporal Data Mining Based on Ontology. *Firt International Conference on Innovative Computing, Infroamtion and Control,* (pp. 535-538). Los Alamitos: IEEE Computer Society.

Yu, S., Aufaure, M. Cullot, N. & Spaccapietra, S.

(2003) Location-Based Spatial Modelling Using Ontology. *Sixth AGILE Conference on Geographic Information Science.* Lyon, France.