Towards the Reduction of Spatial Joins for Knowledge Discovery in Geographic Databases Using Geo-Ontologies and Spatial Integrity Constraints

Vania Bogorny, Paulo M. Engel, Luis O. Alvares

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
{vbogorny, engel, alvares}@inf.ufrgs.br

Abstract. Spatial join is the most expensive operation in geographic databases, but essentially important to compute spatial relationships intrinsic to geographic data. In account of spatial relationships real world entities may affect the behavior of other entities in the neighborhood. Spatial relationships are fundamental for knowledge discovery in geographic databases and are strongly related to the discovered patterns. Knowledge discovery is a user-dependent process, but the user is usually neither an expert in geographic databases nor in spatial relationships. This paper presents an approach to reduce the number of spatial relationships for knowledge discovery, using a geo-ontology and semantic spatial integrity constraints. We show how spatial constraints can help the user of knowledge discovery in both defining the semantically consistent spatial relationships and reducing the verification of unnecessary relationships.

1 Introduction

The increasing use of geographic data in different application domains has resulted in large amounts of data stored in geographic databases. Geographic data are real world entities, also called spatial features [1], which have a location on Earth's surface. All spatial features (e.g. Portugal, 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*). Figure 1 shows an example of spatial feature types, where shape is a spatial attribute characterizing the geometric representation (e.g. point, line or polygon), and the map is a graphical representation of some shapes.

Beyond the spatial attributes, there are implicit spatial relationships, which are intrinsic to geographic data, but usually not explicitly stored in geographic databases (e.g. Roads *cross* Rivers). Because of spatial relationships real world entities can affect the behavior of other features in the neighborhood. These implicit correlations can only be discovered with specific techniques for knowledge discovery.

Knowledge discovery in databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data [2]. In geographic databases knowledge discovery is the extraction of interesting spatial patterns and features, general relationships between spatial and non-spatial data, and other general characteristics of data not explicitly stored in these databases [3].

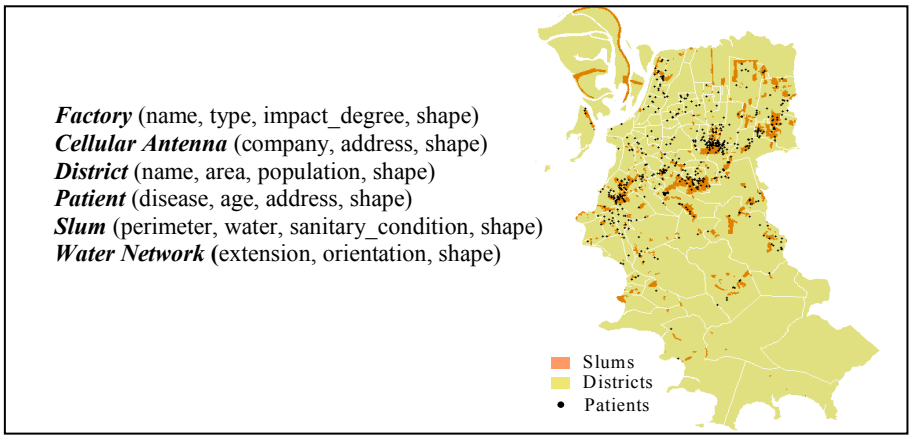

Fig. 1. Spatial feature types and its graphical representation (map)

Spatial join is the operation to compute spatial relationships between two spatial features. It is the most expensive operation in geographic databases for both spatial data analysis and knowledge discovery.

The algorithms for knowledge discovery are not intelligent enough to decide which aspects in geographic databases are relevant or not to the discovery process. The relationships and many other parameters should be provided by the KDD user, what makes the discovery process extremely user-dependent. However, the KDD user is usually not an expert in geographic databases, and he may not have enough background knowledge to decide which aspects to consider in the discovery process.

Geographic data share a large number of spatial relationships, but many are irrelevant to the discovery process and are unnecessarily calculated. For example, an *island* is a piece of land surrounded by *water*. In a geographic database, *island* should be represented as a spatial feature type with a mandatory relationship with a spatial feature type *water*. So, why should we compute spatial relationships between islands and water resources for knowledge discovery if by definition they are related to each other? Why should we consider this kind of relationships if they will create patterns with high confidence without adding novel knowledge? These and other aspects are usually not familiar to the KDD user, but are well-known concepts to geographers or geographic database designers.

Geographic database designers or specialists in Geography know the nature, the concepts, and the semantics of geographic data, so they are able to specify both mandatory and prohibited spatial relationships which define spatial integrity constraints. By specifying these constraints in a geo-ontology, the knowledge of specialists in geographic data can be reused to help the KDD user.

In the literature, there are basically two approaches for knowledge discovery in geographic databases: one is based on quantitative reasoning, which mainly computes distance relationships; and the other is based on qualitative reasoning. Algorithms based on qualitative reasoning [3,4,5,6,7] compute spatial relationships according to a relationships hierarchy, but they neither filter the relationships nor consider if they are geometrically possible or semantically consistent.

In this paper we show how to reduce the number of topological relationships for knowledge discovery in geographic databases with spatial integrity constraints and geo-ontologies. The novelty of our approach is the use of geo-ontologies as prior

knowledge to eliminate mandatory as well as prohibited topological relationships expressed by spatial integrity constraints, and deduce which topological relations may lead to interesting patterns in the KDD process.

The remainder of the paper is organized as follows: Section 2 presents the basic concepts of spatial relationships and spatial constraints. Section 3 presents a geo-ontology meta-model for geographic data and spatial integrity constraints. Section 4 shows how geo-ontologies and spatial integrity constraints can be used as prior knowledge to reduce geographic data pre-processing for knowledge discovery. Finally, Section 5 concludes the paper and suggests some directions of future work.

2 Spatial Relationships and Semantic Integrity Constraints

Geographic data share basically 3 types of spatial relationships: *direction, distance,* and *topological. Direction* relationships deal with the order as spatial features are located in space. *Distance* relations are based on the Euclidean distance between two spatial features. Our focus in this paper is on *topological* relations, which describe concepts of adjacency, containment and intersection between two spatial features.

There are many approaches in the literature to formally define a set of topological relationships among points, lines and polygons [8,9]. The OGC (Open GIS Consortium) [10], which is an organization dedicated for developing 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*.

Considering the geometric representation of spatial features, different topological relationships are applicable. Table 1 shows the topological relationships, standardized by the OGC, considering the **geometry** of two spatial feature types. Empty boxes and checked boxes respectively represent impossible and possible relationships between two geometries. For example, two spatial features represented as line and polygon, respectively, can share the relationships *disjoint, touches, within* and *crosses.*

Table 1. Topological relationships between points, lines and polygons [10]

Topological Relation / Geometric Combination	Disjoint	Overlaps	Touches	Contains	Within	Crosses	Equals
Point(•) Point(•)	✓			✓	✓		✓
Point(•) Line(/)	✓		✓		✓	✓	
Point(•) Polygon(□)	✓		✓		✓	✓	
Line(/) Line(/)	✓	✓	✓	✓	✓	✓	✓
Line(/) Polygon(□)	✓		✓		✓	✓	
Polygon(□) Polygon(□)	✓	✓	✓	✓	✓		✓

Spatial integrity constraints encompass the peculiarities of geographic data and spatial relationships. Their purpose is to warrant as well as to maintain both the quality and the consistency of spatial features in geographic databases. Cockroft [11] proposed three types of spatial integrity constraints: topological, semantic, and user defined constraints. Topological integrity constraints refer to the topological consistency of the shape, such as “the boundary of a state must be contained inside

the shape of the country". Semantic constraints refer to the spatial consistency of spatial features according to their meaning (e.g. "lakes cannot contain rivers"). User defined integrity constraints are equivalent of "business rules" defined in non-geographic databases, such as, "residential areas must lie farther than 1000 meters from a nuclear plant".

Serviane [12] presented topological-semantic integrity constraints, which define mandatory or prohibited topological relationships according to the semantic of the spatial feature. Considering only the geometric representation of spatial features most topological relationships are possible. Considering their meaning, it is possible to define which topological relation is consistent and which one is inconsistent. Extending the approach to specify topological-semantic constraints proposed by Bogorny [13], in order to support the cardinality "all", for mandatory *disjoint* relationships, a topological-semantic constraint between two spatial feature types *A* and *B* can be defined as:

<constraint> ::= <spatialFeatureTypeA><predicate> <spatialFeatureTypeB>
<predicate> ::= <relType> <minCard> <maxCard>
<relType>::='touches'|'overlaps'|'equals'|'within'|'contains'|'crosses'|'disjoint'
<minCard> ::= 0|1| a
<maxCard> ::= 0|1| a

The predicate of a spatial constraint is given by a relationship type *relType,* a minimum cardinality <*minCard*>, and a maximum cardinality <*maxCard*>. The predicate can express mandatory constraints, which are given by the cardinalities (*a,a*) for the relationship *disjoint*, and (1,1) for the remaining topological relationships. A spatial constraint for Hospital with Factory, for example, can be defined as <Hospital> <*disjoint*><*a*><*a*><Factory>, where all instances of Hospital are *disjoint* to ALL instances of Factory. A spatial constraint for Island with Water Resource, for example, where every Island has a *within* relationship with only one Water Resource can be expressed such as: <Island> <*within*> <1><1> <Water Resource>.

Prohibited constraints are defined through the cardinalities (0,0). For example, <River> <*contains*> <0><0> <Road>.

3 Geo-Ontologies

Ontology is an explicit specification of a conceptualization [14]. 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 [15]. It is an agreement of both the concepts meaning and the structure of a specific domain. Each concept definition must be unique, clear, complete, and non-ambiguous. The structure represents the properties of the concept, including a description, attributes and relationships with others concepts.

Ontologies have been used recently in many and different fields in Computer Science, such as Artificial Intelligence, Databases, Conceptual Modeling, Semantic Web, etc. Although research is not so far yet in ontologies for geographic data [16], some geo-ontologies have been emerging recently. Besides defining a geo-ontology for administrative data for the country of Portugal, Chaves [17] defines a geo-ontology meta-model, named GKB (Geographic Knowledge Base).

GKB provides the concept of spatial *Feature,* which is represented as a class, and is associated to a *Feature_Type*, whose instances represent all feature types specified for a domain. For example, Country is an instance of *Feature_Type,* while Brazil and Portugal are instances of *Feature*. The class *Name* has names identified for every feature in all available information sources, including synonyms. Concepts of relationships among features in GKB are specified through the classes *Relationship* and *Relationship_Type,* which can assume concepts of *partOf* and *adjacency*.

In our point of view, a geo-ontology should provide the definition of the main aspects of geographic data, which are already defined in geographic meta-models for conceptual modeling (e.g. MADS, OMT-G) and standardized by the OGC. Based on these definitions, a geographic concept should have, at least: one spatial attribute given by a geometry, non-spatial attributes, relationships with other geographic concepts, and spatial constraints. The relationships can be conventional, such as aggregations or associations, or spatial, such as topological, distance or order. Considering these characteristics, we extended the GKB proposed in [17] to support geometry and spatial integrity constraints.

Figure 3 shows the extended GKB meta-model. The classes *GM_Object* and *GM_ObjectType* were added following the OGC definitions. *GM_ObjectType* represents the geometric representation of a feature type (e.g. point, line, and polygon). *GM_Object* is an instance of a geometric type associated to a specific feature. The cardinalities 0, 1, and *a* added to the dual relationship between the classes *Relationship* and *Feature* define concepts of mandatory or prohibited constraints.

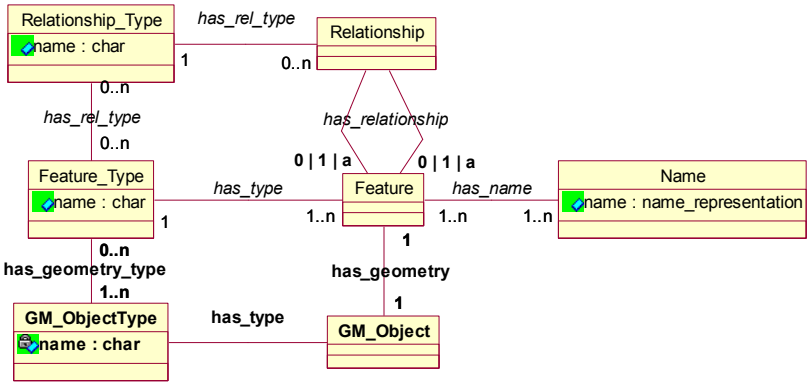

Fig. 3. Extended GKB to support geometry, topological relationships and spatial constraints

4 Geo-Ontologies and the KDD Process

The possible binary topological relationships between two geometric objects shown in Table 1 can be significantly reduced if we consider the semantics of each object. Table 2 shows an example of the same geometric combinations illustrated in Table 1, giving a different semantics to each geometric object. The geometries point and line, for example, can share the relationships *disjoint, touches, within* and *crosses* (see Table 1). Considering that point and line have respectively the semantics of Bridge

and River (see Table 2), then only *crosses* is semantically consistent. The combinations line/line, for example, can share any topological relation, but if their semantics is respectively River and Road, then only *disjoint, touches* and *crosses* are consistent (see Table 2). For the combination polygon/polygon with the semantics State and Country respectively, only the relationship *within* is consistent.

Table 2. Possible topological relationships considering the semantics of the feature types

Topological Relation / Semantic Combinations	Disjoint	Overlaps	Touches	Contains	Within	Crosses	Equals
Factory (●) Hospital (●)	✓						
Bridge (●) River (/)						✓	
Factory (●) Airport(□)	✓		✓				
River (/) Road (/)	✓		✓			✓	
Beach (/) Sea (□)			✓				
State (□) Country (□)					✓		

Although the topological relationships shown in Table 2 are semantically possible, not all of them are interesting for knowledge discovery. So, if beside considering the semantics of the features we also consider spatial integrity constraints, it is possible to reduce still more the number of spatial joins and define which relationships should be computed for knowledge discovery.

Applying spatial integrity constraints, Table 3 shows the possible topological relationships between the same feature types shown in Table 2. Considering only the semantics of the spatial feature types, we would have 9 possible relationships according to the example shown in Table 2. Considering spatial integrity constraints we would have only 3 relevant relationships to consider in the discovery process.

Table 3. Topological relationships for knowledge discovery

Topological Relation / Semantic Combinations	Disjoint	Overlaps	Touches	Contains	Within	Crosses	Equals
Factory (●) Hospital (●)							
Bridge (●) River (/)							
Factory (●) Airport(□)			✓				
River (/) Road (/)			✓			✓	
Beach (/) Sea (□)							
State (□) Country (□)							

On the one hand, the prohibited constraints forbid the inconsistent relationships, so they should not exist in the database. By consequence, they do not need to be computed for spatial analysis or knowledge discovery. On the other hand, mandatory relationships will produce patterns with high confidence in the discovery process because mandatory relationships will always hold if the database is consistent. However, these patterns will not add novel knowledge to the discovery.

Despite mandatory and prohibited constraints do not explicitly define the relevant relationships for knowledge discovery, we are able to eliminate those which are

mandatory or prohibited, and specify those which are possible. Let us consider the set of all topological relationships as R = {*touches, contains, within, crosses, overlaps, equals, disjoint*}. T is the set of topological relationships **geometrically** possible between two feature types A and B. Pr is the set of **prohibited** relationships between A and B, M is the set of **mandatory** relationships and P_{KDD} is the set of **possible** relationships for knowledge discovery. If a prohibited constraint is given between A and B, then the set of possible relationships is $P_{KDD(A,B)} = T_{(A,B)} - Pr_{(A,B)}$. If a mandatory constraint is defined between A and B, then $P_{KDD(A,B)} = \phi$.

The approximate reduction cost of computing spatial joins for each pair of spatial feature types A and B for knowledge discovery is given by $R_{cost(A,B)} = (|T_{(A,B)}| - |P_{KDD(A,B)}|)$. $Cost_{re(A,B)}$, where $Cost_{re\ (A,B)}$ is the time to compute each topological relationship between A and B. The cost to browse the geo-ontology is not considered.

In the discovery process, a data pre-processing algorithm can compute the topological relationships according to the properties of the feature types specified in the geo-ontology. For example, let us consider that the feature type of interest specified by the KDD user is *River* and that the relevant feature types to be spatially compared with *River* are *Road*, *Hospital*, and *Island*. Suppose that in a geo-ontology *River* has the properties of a mandatory relationship *disjoint* with *Hospital* and a prohibited relationship *contains*, *overlaps*, *inside* and *equals* with *Road*, but no property with *Island*. The first step of the pre-processing algorithm is to read the properties of *River* and specify that $P_{KDD(River,Road)}$ = {*touches, crosses*} and $P_{KDD(River,Hospital)} = \phi$. As P_{KDD} is already defined for *Road* and *Hospital*, the second step is to read the properties of *Island* in the geo-ontology, and specify $P_{KDD(River,Island)}$. Suppose that *Island* has the property of a mandatory relationship *within*, with *River*, than $P_{KDD(River,Island)} = \phi$.

5 Conclusions and Future Work

In this paper we presented a geo-ontology meta-model to define concepts and properties of geographic data. Through the properties we can specify spatial integrity constraints, which forbid or obligate specific topological relationships between specific feature types.

Considering only the geometry of spatial feature types, a certain number of topological relationships is possible. We showed how this number can be reduced if we consider the semantics of the spatial features and their spatial integrity constraints, using geo-ontologies. We also showed how the spatial integrity constraints can contribute for knowledge discovery in geographic databases. The mandatory and the prohibited spatial relationships defined by the constraints are irrelevant to the discovery process because of two reasons: - *prohibited relationships* will never exist if the database is consistent; and - *mandatory relationships* will produce patterns with high confidence but which do not add any novel knowledge to the discovery process.

As future work, we will study the application of distance and order constraints and how we can reduce the number of spatial joins for the KDD process with different combinations of spatial relationships.

6 ACKNOWLEDGMENTS

We would like to thank CAPES and CNPQ for the financial support of this research, and Stefano Spaccapietra and Daniela Leal Musa, for their comments.

7 References

1. Opengis. The Opengis Abstract Specification Topic 1: Feature Geometry. In URL: Http://Www.Opengeospatial.Org/Docs/01-101.Pdf (2001).
2. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P.: Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, Portland, OR, (1996) 82-88.
3. Lu, W., Han, J., Ooi, B. C.: Discovery of General Knowledge In Large Spatial Databases. Workshop on Geographic Information Systems, Singapore (1993) 275-289.
4. Ester, M., Frommelt, A., Kriegel, H.-P., Sander, J.: Spatial Data Mining: Database Primitives, Algorithms And Efficient DBMS Support. Journal of Data Mining and Knowledge Discovery. 4 (2000) 193-216.
5. Malerba, D., Appice, A., Vacca N.: SDMOQL: An OQL-Based Data Mining Query Language For Map Interpretation Tasks. In: Workshop On Database Technologies For Data Mining. Springer, Prague, Czech Republic (2002).
6. Appice, A., Ceci, M., Lanza, A.: Discovery of Spatial Association Rules In Geo-Referenced Census Data: A Relational Mining Approach. Intelligent Data Analysis. Software & Data 6 (2003).
7. Koperski, K., Han, J.: Discovery of Spatial Association Rules In Geographic Information Databases. In: International Symposium In Large Spatial Databases. Springer, Portland, Maine, USA (1995) 47-66.
8. Clementini, E., Di Felice, P.: A Model for Representing Topological Relationships between Complex Geometric Features In Geographical Databases. Information Sciences. 90 (1996) 121-136.
9. Egenhofer, M., Herring, J.: Categorizing Binary Topological Relations Between Regions, Lines, and Points In Geographic Databases. Technical Report TR-941, University of Maine, (1994).
10. Opengis. Open GIS Simple Features Specification For SQL. In URL: Http://Www.Opengeospatial.Org/Docs/99-054.Pdf (1999).
11. Cockcroft, S.: A Taxonomy of Spatial Data Integrity Constraints. Geoinformatica, Kluwer Academic Publishers, Hingham, MA, USA. 1-4 (1997) 327-343.
12. Servigne, S. et al.: A Methodology for Spatial Consistency Improvement of Geographic Databases. Geoinformatica. 4-1 (2000) 7-34.
13. Bogorny, V., Iochpe, C.: Extending the OpenGIS Model to Support Topological Integrity Constraints. In: Brazilian Symposium on Databases, COPPE/UFRJ, Rio de Janeiro, Brazil (2001) 25-39 (in Portuguese).
14. Gruber, T. R.: Towards Principles for the Design of Ontologies Used for Knowledge Sharing. Formal Ontology in Conceptual Analysis and Knowledge Representation. Int. J. of Human-Computer Studies. Kluwer Academic Publishers. 43 (1993) 907-928.
15. Guarino, N.: Formal Ontology and Information Systems. In: International Conference on Formal Ontology in Information Systems. Italy (1998) 3-15.
16. Spaccapietra, S., Cullot, N., Parent, C., Vangenot, C.: On Spatial Ontologies. In: Brazilian Symposium on GeoInformatics. Campos do Jordão, Brazil. (2004).
17. Chaves, M. S., Silva, M. J., Martins, B.: A Geographic Knowledge Base for SemanticWeb Applications. In Brazilian Symposium on Databases, Uberlandia, Minas Gerais, Brazil (2005). To appear.