

Enhancing Spatial Association Rule Mining in Geographic Databases

Vania Bogorny^{1,2}, Paulo Martins Engel², Luis O. Alvares²

¹ Hasselt University & Transnational University of Limburg
Agoralaan Gebouw D, 3590 Diepenbeek - Belgium

² Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

vania.bogorny@uhasselt.be, {engel, alvares}@inf.ufrgs.br

Abstract. Association rule mining algorithms generate huge amounts of patterns. In spatial association rule mining this problem increases because a significant amount of associations is well known a priori. This paper presents a novel approach for mining spatial association rules, using background knowledge. The main contributions include the use of geographic database schemas and geo-ontologies for (i) the improvement of geographic data pre-processing, (ii) the elimination of well known patterns and (iii) the generation of maximal frequent patterns without redundant and non-interesting associations.

1. Introduction

The association rule mining technique [Agrawal and Srikant 1994] has been largely used for knowledge discovery in databases. As a data mining task it emerged with the objective to find novel, useful, and interesting patterns, *hidden* among data. An enormous amount of algorithms has been proposed for mining association rules, but their main drawback is the generation of huge amounts of rules. To reduce this problem different objective and subjective measures have been proposed, but only the data without the schema has been considered so far.

In transactional association rule mining items and transactions are normally stored in one single relation, and the schema may not be useful. In geographic databases, however, different object types (e.g. streets, water resources) are stored in different relations [Gutting 1994]. From the database design point of view the objective of data modeling is to bring together all relevant object types of the application, their relationships, and their constraints [Elmasri and Navathe 2003], [Shekhar and Chawla 2003]. Many geographic object types have mandatory associations, which we call *geographic dependences*, usually represented in the third normal form, intending to reduce anomalies and warrant integrity.

In contrast to database schema modeling, where associations between data are *explicitly* represented, association rule mining algorithms should find *implicit* and *novel* associations. While the former represents the data into the third normal form, the later usually denormalizes the data in one single table. This transformation brings together associations explicitly represented in the schema to the dataset to be mined, and by consequence many well known associations are extracted by rule mining algorithms and presented to the user.

Geographic dependence is a mandatory *one-one* or *one-many* relationship between two geographic object types *A* and *B* where each instance of *A* is related to at least one instance of *B*. The number of associations specified in geographic database schemas reflects a large number of well known geographic dependences, as shown with two case studies in [Bogorny et al. 2006c]. However, spatial relationships (see [Gutting 1994] for details) are the main characteristic which differs geographic data from transactional data, and therefore need to be considered for data

mining [Koperski and Han 1995]. Figure 1(a) shows a well known relationship where all gas stations intersect streets. This pattern is represented in the respective schema in Figure 1(c) by the one-many relationship of gas station with street. If considered in spatial association rule (SAR) mining such dependence will produce meaningless rules such as $is_a(GasStation) \rightarrow intersect(Street)$. In Figure 1(b) there is no obvious pattern between gas stations and water resources, and in the schema in Figure 1(c) there is no relationship between this pair. Therefore, their spatial relationships might be interesting for knowledge discovery.

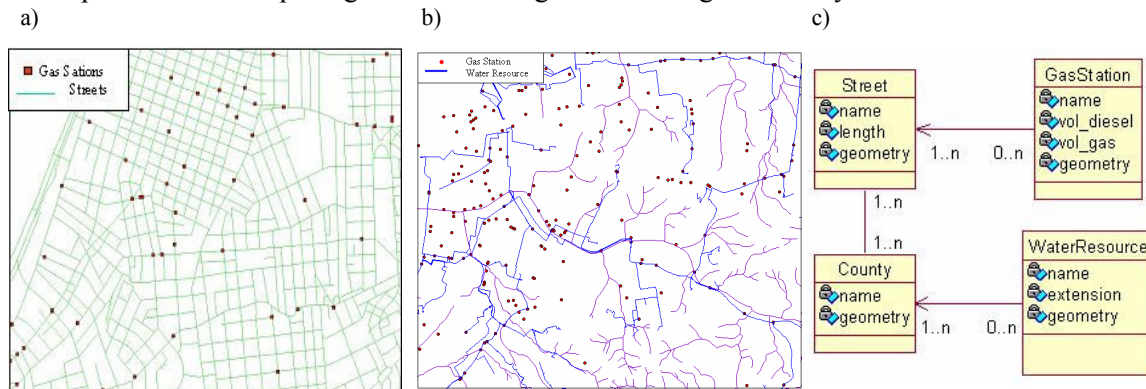


Figure 1. (a) Well known relationships, (b) non-standard relationships, and (c) schema.

Geographic data may generate hundreds or thousands of interesting and uninteresting associations that can discourage users from interpreting them in order to find interesting patterns. We claim that well known associations should be eliminated in SAR mining to avoid their extraction and presentation to the user. For this purpose we propose new methods for mining SAR using background knowledge.

The remaining of this paper is organized as follows: Section 2 presents an overview of how well known dependences can occur in SAR mining. In Section 3 we discuss the related works and what has been done so far to reduce the number of well known SAR. In Section 4 we present our approach and the main contributions, and in Section 5 we conclude the paper.

2. The Main Problem of Geographic Dependences in SAR Mining

A spatial association rule is an implication of the form $X \rightarrow Y$ with support s and confidence c , where at least one element in X or Y is a spatial predicate [Koperski and Han 1995]. In SAR mining the first task is the computation of spatial predicates (relationships) followed by the generation of frequent sets and association rules. Well known geographic dependences can appear in both steps, as will be explained in the following sections.

2.1. Geographic Dependences in Spatial Relationship Computation

In transactional data mining, every row in the dataset to be mined is a transaction and columns are items, while in spatial data mining every row is normally an instance of a reference object type (e.g. city), called *target feature type*, and columns are predicates. Every predicate is related to a non-spatial attribute (e.g. name) of the target feature type, or a *relevant feature type* that is spatially related to the target feature type (e.g. $contains(gasStation)$). Table 1 shows a preprocessed spatial dataset where every row is a city and the predicates are relevant geographic object types (port, water body, hospital, treated water network, factory) spatially related to city.

Spatial predicates are materialized spatial relationships extracted with spatial joins between all instances (e.g. Porto Alegre) of a target feature type T (e.g. city) and all instances (e.g. Ipiranga) of every relevant feature type O (e.g. street) in a set of relevant feature types S (e.g. street, hospital). This process is the computational bottleneck in spatial data mining.

Well known geographic dependences may exist between T and O , or between pairs of $O \subseteq S$. In Table 1 there is a well known dependence between the target feature type city and treated water network, because every city has at least one treated water network. This means that the predicate *contains(TreatedWaterNet)* has a 100% support and a large number of rules with this predicate will be generated, such as *contains(factory) → contains(TreatedWaterNet)*. Such a rule expresses that cities that contain factories do also contain treated water networks. Although the rule seems to be interesting, it can be considered obvious due the simple fact that *all* cities have treated water networks, having they factories, or not.

Table 1. Preprocessed spatial dataset for mining SAR

City	Predicates
1	contains(Port), contains(Hospital), contains(TreatedWaterNet), contains(Factory), crosses(WaterBody)
2	contains(Hospital), contains(TreatedWaterNet), crosses(WaterBody)
3	contains(Port), contains(TreatedWaterNet), contains(Factory), crosses(WaterBody)
4	contains(Port), contains(Hospital), contains(TreatedWaterNet), crosses(WaterBody)
5	contains(Port), contains(Hospital), contains(TreatedWaterNet), contains(Factory), crosses(WaterBody)
6	contains(Hospital), contains(TreatedWaterNet), contains(Factory)

Predicates with 100% support appear in most generated rules as can be observed in Table 2, which shows the result of an experiment performed over Table 1. For minimum support 20% 31 frequent sets and 180 rules were generated using the algorithm Apriori [Agrawal and Srikant 1994]. Among the 31 frequent sets and the 180 rules, 16 frequent sets and 130 rules had the dependence *contains(TreatedWaterNet)*. 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 had the dependence.

Table 2. Frequent sets and association rules generated with Apriori

MinSup %	All Frequent Sets/ Rules	FrequentSets with dependence / FrequentSets without dependence	Rules with Dependence / Rules without Dependence
20	31 / 180	16/15	130/ 50
50	25 / 96	13/12	72 / 24

In this section we showed the number of frequent patterns that can be generated by a dependence between the target feature type and a relevant feature type. These dependences can be removed by pruning the input space, as will be explained in Section 4. However, dependences may exist among relevant feature types. In Table 1 *Port* has a dependence with *Water Body*, since all cities which have *Ports* must also have *Water Bodies*. This kind of dependence cannot be removed by pruning the input space, as will be explained in the following section.

2.2. Geographic Dependences in Frequent Sets

Two main tasks are related to the extraction of association rules [Agrawal and Srikant 1994]: (i) *Find all frequent predicates sets*: a set of predicates is frequent if its support is at least equal to a certain threshold, called *minsup*, and (ii) *Generate strong rules*: a rule is strong if its support is at least equal to *minsup* and the confidence is at least equal to a threshold called *minconf*.

To find *frequent predicate sets* and *extract association rules*, relevant feature types O in S are combined with each other for the different instances of the target feature type T , and not among T and O as in the previous problem. To illustrate the geographic dependence replication process in the frequent set generation, Table 3 shows the frequent sets extracted from the dataset in Table 1 with minimum support 50%, where k is the number of elements in the frequent sets and the dependence is the pair $\{contains(Port), crosses(WaterBody)\}$.

A geographic dependence appears the first time in frequent sets with $k=2$ elements, and can be replicated to many frequent sets of size $k>2$ when minimum support is reached, as shown in bold style in Table 3. Considering such a small example and high minimum support (50%),

one single dependence participates in 6 frequent sets, which represents 30% of the total number of frequent sets. The number of rules with a geographic dependence will be much larger than the frequent sets, mainly when the dependence is frequent in the largest frequent set.

Table 3. Frequent predicate sets with minimum support 50%

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

Geographic dependences among relevant feature types cannot be removed by pruning the input space. In our example, either *Water Body* or *Port* may have an interesting association with any other relevant feature type.

3. Related Works and Contribution

There are two main approaches in the literature for mining SAR. The first is based on quantitative reasoning [Yoo et al. 2004, 2005], which computes distance relationships during the frequent set generation. It has the advantage of not requiring the definition of a reference object, but has some general drawbacks such as deal only with points and do not consider non-spatial attributes. For spatial objects represented by lines or polygons, their centroid is extracted. This process may loose information and generate non-real patterns (e.g. the Mississippi River *intersects* many states as a multi-line object, but is *far from* the same states by considering its centroid).

The second approach is based on qualitative spatial reasoning [Appice et al. 2005][Mennis et al. 2005][Koperski and Han 1995] and considers distance and topological relationships between a reference geographic object type and a set of relevant feature types represented by any geometric primitive (e.g. points, lines, and polygons). In [Koperski and Han 1995] and [Mennis et al. 2005] spatial association rules are pruned using minimum support and confidence, but rule interestingness is not addressed. In [Appice et al. 2005] all spatial relationships are extracted from the database and transformed into a deductive relational database, which is non-trivial for real applications. Well known rules are partially pruned with declarative bias defined by the user, but in post processing steps, after the generation of frequent sets and association rules.

Our framework follows the qualitative spatial reasoning approach and removes well known patterns completely by early pruning the input space and the frequent itemsets. Indeed, it is applicable to any algorithm that generates frequent sets, and has as the main advantage its simplicity. While dozens of approaches for mining non-spatial association rules define syntactic constraints and do later pruning, we consider *knowledge constraints*, and eliminate the exact well known patterns which generate non-interesting association rules. For instance, a pair {*gas station*, *street*} is a knowledge constraint used as background knowledge in the data mining step.

4. The Proposed Framework

To illustrate the different contributions of this work, Figure 2 shows an overview of our framework for enhancing SAR mining in geographic databases. This framework supports the complete discovery process [Fayyad et al. 1995] and can be analyzed at three general levels: data repository, data preprocessing, and data mining.

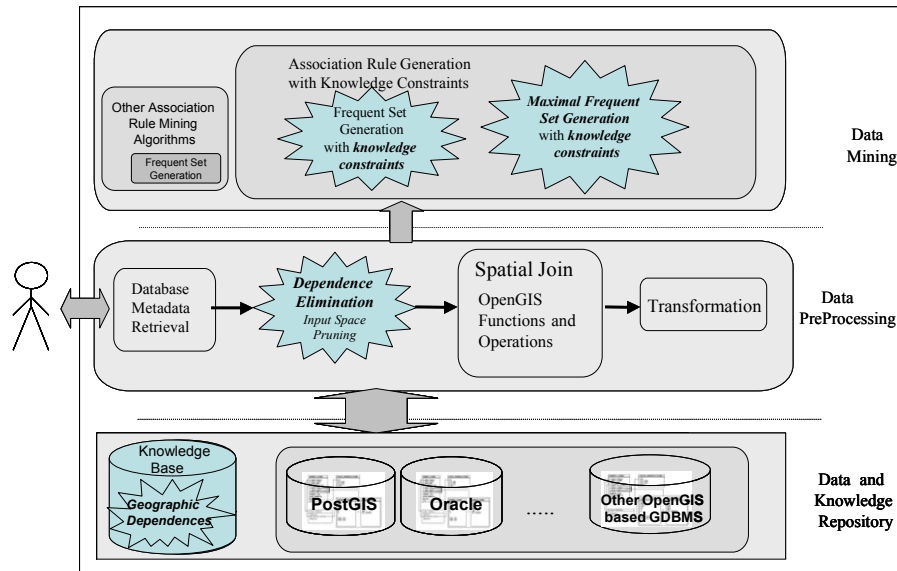


Figure 2. GeoARM+: a Unified Framework for mining SAR in geographic databases.

On the bottom are the geographic databases, stored in geographic database management systems constructed under OGC specifications [OGC 1999]. There is also a knowledge base which stores information that can be used as background knowledge. This repository may store geographic database schemas, geo-ontologies, concept hierarchies, or a set of pairs of geographic objects with dependences defined by the user. On this level we developed two algorithms to automatically extract geographic dependences, also called *knowledge constraints*, from geographic database schemas [Bogorny et al. 2006b] and geo-ontologies [Bogorny et al. 2007a]. In the center of Figure 2 is the spatial data preprocessing level which covers the *gap* between data mining tools and geographic databases. At this level data repositories are accessed and data are retrieved, preprocessed, and transformed into the single table format. An overview of our contributions at this level is given in section 4.1. On the top are the algorithms for mining SAR, and our contributions to this level will be explained in Section 4.2.

4.1. Data Preprocessing Tasks: The Input Space Pruning Method

There are four main steps to perform the tasks of extracting spatial predicates for mining SAR: *database metadata retrieval*, *dependence elimination (input space pruning)*, *spatial join*, and *transformation*, as shown in Figure 2. The *database metadata retrieval* step connects to the specified database and retrieves relevant information through the Open GIS database schema metadata, including all database relations (feature types). The user can choose among the database relations the target feature type T and its non-spatial attributes A , the set S of relevant feature types O that may have some influence on T , the spatial relationships, and a granularity level g for each different relevant feature type $O \subseteq S$ when a concept hierarchy h_O is given for O [Bogorny 2006f].

The *Dependence Elimination* step verifies all associations between T and S . It searches the set of knowledge constraints and if T has a dependence with any $O \subseteq S$, then O is eliminated from S . For each relevant feature type removed from S , no spatial join is required to extract

spatial relationships. By consequence, neither frequent sets nor SAR will be generated with this relevant feature type. This step we call *input space pruning*, which according to [Bonchi et al. 2003] is the most efficient way to reduce frequent sets and association rules. The efficiency and effectiveness of this method was presented in [Bogorny et al. 2006a] using geographic database schemas and the algorithm Apriori [Agrawal and Srikant 1994].

The *Spatial Join* step computes and materializes the spatial relationships between T and O , for all $O \subseteq S$ at the granularity level g_O , according to a concept hierarchy h_O . Details of this process can be found in [Bogorny 2006f]. If no concept hierarchies are provided, two different granularity levels can be automatically generated, as explained in [Bogorny 2005a]. Four types of spatial relationships can be extracted: *topological*, *intersection*, *order*, and *distance*. Since this is the most time consuming step in SAR mining we proposed to use geo-ontologies to optimize the extraction of spatial relationships [Bogorny et al. 2005b, 2007a] and to partially remove well known dependences [Bogorny et al. 2007a]. This method was evaluated with the real geographic ontology named geo-netpt01 of the country of Portugal. Spatial joins are computed with spatial operations following the Open GIS Consortium specifications [OGC 1999], what makes the proposed framework interoperable with all GIS constructed under these standards.

The *Transformation* step transposes as well as discretizes the dataset Ψ into the single table representation understandable by most association rule mining algorithms.

Since to the best of our knowledge there is no available spatial data mining toolkit, the data preprocessing steps here presented resulted in an extension of the open source data mining toolkit Weka [Witten and Frank 2005] to support automatic geographic data preprocessing [Bogorny et al. 2006e] and partial dependence elimination. In this system the user can automatically generate spatial datasets at two different granularity levels as well as specify well known dependences.

4.2. New Frequent Pattern Mining Methods: Apriori-KC and Max-FGP

The most efficient way to eliminate well known geographic dependences in SAR mining that cannot be eliminated in data preprocessing is to eliminate candidate sets in which the dependences appear the first time, and before extract association rules. For this purpose we proposed the Apriori-KC [Bogorny et al. 2006b] algorithm to generate frequent geographic patterns from geographic database. In this method we exploit the anti-monotone [Bonchi et al. 2003] constraint and eliminate well known geographic dependences among relevant feature types during the candidate generation. However the elimination of dependences among relevant feature types is not as effective as the elimination of dependences between the target feature type and the relevant feature types. Therefore, in [Bogorny et al. 2006c] we proposed to prune the input space as much as possible in one step and the frequent sets in a second step. In [Bogorny et al. 2006d] we proposed Max-FGP, which is a novel method for mining maximal generalized frequent geographic patterns in which both redundant and non-interesting patterns are removed. We applied the closed frequent pattern mining technique [Bastide et al. 2000] which has not been applied to the geographic domain so far. The algorithm Max-FGP reduced the number of frequent patterns in relation to Apriori-KC in more than 70% in all our experiments [Bogorny 2006f], which are omitted because of space limitations.

Besides the non-interesting patterns generated by geographic dependences other different kinds of non-interesting patterns are extracted from geographic data [Bogorny 2006f]. Large amounts of rules are generated with the same geographic object type and different spatial relationships such as *contains(Slum) \rightarrow touches(Slum)*. To remove such associations we proposed Apriori-KC⁺ [Bogorny et al. 2007b], which is an extension of Apriori-KC [Bogorny et al. 2006b].

A summary of our contributions is presented in the IJGIS international journal [Bogorny 2007c] and in [Rinzivillo 2007d].

5. Conclusions and Future Works

In this work we presented an approach to reduce the high number of SAR extracted from geographic databases. We presented contributions to the whole process of mining SAR, not only to one specific algorithm. Not only is the number of association rules reduced, but rules which represent well known patterns are eliminated as well. While existing methods for non-spatial data propose objective and subjective measures to reduce the number of rules, we use background knowledge to eliminate the exact pairs of geographic objects which generate non-interesting patterns. The main strength of our framework is its simplicity. Although in this work we have focused on geographic data, the proposed methods are application independent, since non-interesting patterns can be extracted in relational databases of different domains, and it is still a challenge to show only non-redundant and interesting patterns to the user.

The thesis summarized in this paper resulted in a significant number of scientific publications which are mentioned into the paper. Additional results of this work include: a post-doc on the context of the European Project GeoPKDD (<http://www.geopkdd.eu>); a Master dissertation at the Hasselt University co-supervised by Vania Bogorny; and a Master dissertation as well as a bachelor work at UFRGS supervised by professor Luis Otavio Alvares.

ACKNOWLEDGMENT

This research has been partially funded by the Brazilian agencies CAPES and CNPq, and the European Union (FP6-IST-FET programme, Project n. FP6-14915, GeoPKDD: Geographic Privacy-Aware Knowledge Discovery and Delivery, (www.geopkdd.eu)).

References

- Agrawal, R. and Srikant, R. (1994) "Fast Algorithms for Mining Association Rules in Large Databases". In *20th VLDB*, Chile, p.487–499. Morgan Kaufmann.
- Appice, M., Berardi, M., Ceci, M. and Malerba, D. (2005) "Mining and Filtering Multi-level Spatial Association Rules with ARES". In *15th ISMIS*, New York, p.342-353. Springer.
- Bastide, Y., Pasquier, N, Taouil, R, Stumme, G, and Lakhal, L (2000). "Mining Minimal Non-Redundant Association Rules using Frequent Closed Itemsets". In *1st CL*, London, p. 972-986.
- Bogorny, V., Engel, P.M. and Alvares, L.O.(2005a) "A Reuse-Based Spatial Data Preparation Framework for Data Mining". In *17th SEKE*, Taiwan, p.649-652. Knowledge Systems Institute.
- Bogorny, V.; Engel, P. M. and Alvares, L.O. (2005b) "Towards the Reduction of Spatial Joins for Knowledge Discovery in Geographic Databases using Geo-Ontologies and Spatial Integrity Constraints". In *2nd KDO ECML/PKDD Workshop*, Porto. p.51-58.
- Bogorny, V., Engel, P. M. and Alvares, L.O (2006a) "GeoARM: an Interoperable Framework to Improve Geographic Data Preprocessing and Spatial Association Rule Mining". In *18th SEKE*, San Francisco, p. 79-84. 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-IS*, London, p.532-537. IEEE Computer Society.
- Bogorny, V.; Camargo, S.; Engel, P.; Alvares, L. O. (2006c). "Mining Frequent Geographic Patterns with Knowledge Constraints". In *14th ACM-GIS*, Arlington, p.139-146. ACM Press.

- Bogorny, V.; Valiati, J.; Camargo, S.; Engel, P.; Kuijpers, B.; Alvares, L. O. (2006d). "Mining Maximal Generalized Frequent Geographic Patterns with Knowledge Constraints". In *6th IEEE-ICDM*, Hong-Kong, p.813-817. IEEE Computer Society.
- Bogorny, V.; Palma, A; Engel, P.; Alvares, L.O. (2006e). "Weka-GDPM: Integrating Classical Data Mining Toolkit to Geographic Information Systems". In: *2nd WAAMD SBBD Workshop*, Florianopolis, p. 9-16.
- Bogorny, V. (2006f). Enchancing Spatial Association Rule Mining in Geographic Databases. PhD thesis. UFRGS.
- Bogorny, V.; Engel, P. M.; Alvares, L.O.(2007a). "Enhancing the Process of Knowledge Discovery in Geographic Databases Using Geo-Ontologies". In *Data Mining with Ontologies: Implementations, Findings, and Frameworks*, Idea Group, Pennsylvania. (to appear).
- Bogorny, V.; Moelans, B.; Alvares, L.O. (2007b). "Filtering Frequent Geographic Patterns with Qualitative Spatial Reasoning". In *1st STDM IEEE-ICDE Workshop*, Turkey. (to appear).
- Bogorny, V., Kuijpers, B., Alvares, L.O.(2007c). Reducing Non-Interesting Spatial Association Rules in Geographic Databases using Background Knowledge: a Summary of Results. *IJGIS International Journal of Geographical Information Science*, Taylor and Francis. (to appear).
- Bonchi F., Giannotti F, Mazzanti A. and Pedreschi D (2003). "ExAMiner: Optimized Level-wise Frequent Pattern Mining with Monotone Constraints". In *4th IEEE-ICDM*, Florida, p.11-18, IEEE Computer Society.
- Elmasri, R. and Navathe, S. (2003) Fundamentals of database systems, Addison Wesley. 4th Edition.
- Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) From data mining to discovery knowledge in databases. *AI Magazine*, 3(17), 37-54.
- Gutting, R. H. (1994) An Introduction to Spatial Database Systems. *International Journal on Very Large Data Bases*, v3 (4), (October), p. 357 - 399.
- Koperski, K. and Han, J. (1995) "Discovery of spatial association rules in geographic information databases," In *4th SSD*, Portland, p. 47-66. Springer.
- Mennis, J. and Liu, J.W. (2005) Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change. *Transactions in GIS*, v9 (1), (January), p. 5-17.
- OGC (1999). "OpenGIS simple features specification for SQL". <http://www.opengeogeographic.org/docs/99-054.pdf>, accessed August 2005.
- Rinzivillo, S., Turini, F., Bogorny, V., Koerner, C., Kuijpers, B. and May, M. (2007d) "Knowledge Discovery from Geographical Data", In *Mobility, Privacy, Geography: a knowledge discovery perspective*, Springer. (to appear).
- Shekhar, S. and Chawla, S. (2003). Spatial databases: a tour. Prentice Hall. 1st Edition.
- Witten, I. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann. 2nd Edition.
- Yoo J.S., Shekhar S. and Celik M. (2005). "A Join-less Approach for Co-location Pattern Mining: A Summary of Results", In *5th IEEE-ICDM*, Houston, 2005, p.813-816. IEEE Computer Society.
- Yoo, J.S. and Shekhar S. (2004). "A partial join approach for mining co-location patterns". In *12th ACM-GIS*, Washington, p.241-249, ACM Press.