[AM-08-068] Rule Learning for Spatial Data Mining

Abstract

Recent research has identified rule learning as a promising technique for geographic pattern mining and knowledge discovery to make sense of the big spatial data avalanche (Koperski & Han, 1995; Shekhar et al., 2003). Rules conveying associative implications regarding locations, as well as semantic and spatial characteristics of analyzed spatial features, are especially of interest. This overview considers fundamentals and recent advancements in two approaches applied on spatial data: spatial association rule learning and co-location rule learning.

Keywords: geocomputation, pattern recognition, rule learning, spatial association, spatial data mining

Author & citation

Dao, T.H.D. (2018). Rule Learning for Spatial Data Mining. The Geographic Information Science & Technology Body of Knowledge (1st Quarter 2018 Edition), John P. Wilson (ed.). DOI: 10.22224/gistbok/2018.1.3

This Topic is also available in the following editions:

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Rule learning. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers.

Explanation

1. Definitons

unit of learning (UoL): the UoL is the major entity those characteristics or attributes are learned.

data input: data for rule learning is typically in the format of a table (singular relational format) or multiple related tables (multiple relational format). The rows in the table(s) are records that represent the instances of the UoL. The columns are the UoL's attributes which take on values in various formats (e.g., Boolean, numeric, categories).

predicates: predicates are the basic building blocks of rules, used to express an attribute-value test for the UoL. Predicates can be a simple test (e.g., population = large) or a test of a more complex relationship (e.g., close-to “high_income_neighborhood”). The set of all predicates, P, is constructed from the data records, attributes, and attribute values.

Spatial predicates are distinguished from non-spatial (semantic) predicates, as they contain spatial attributes. These attributes relate to the spatial feature characteristics (i.e. location and shape) or the spatial relations among the features. The predicate close-to “high-income-neighborhood” mentioned above is spatial as it is distance-based.

a rule: a rule is an implication typically taking on the form of IF 'conditions' THEN 'result' expression. The conditions part is called the body (or the antecedent), while the result part is called the head (or the consequent) of the rule. The building blocks for the conditions part and the result part of the rule are predicates. The conditions part may be a conjunctive condition connecting various predicates via the AND Boolean operator. The length of a rule is the total number of predicates it contains.

a spatial rule: a spatial rule is a rule containing at least one spatial predicate.

2. Rule Learning in Data Mining

Rule learning, in general, refers to the process of extracting rules expressing regularities (i.e., patterns) found in the data. Resulting rules are then used in knowledge discovery, and in various complex soft computing modelling approaches (e.g., neural networks (NN) and agent-based modelling (ABM)).

Classical rule learning in data mining was originally designed to mine patterns in relational data. In this case, the unit of learning (UoL) includes records with aspatial attributes or shopping transactions. Rules describing co-occurrences of attribute values or of shopping items (expressed as predicates) are of interest. Classical rule learning problems are broadly categorized into descriptive and predictive, although the separation is not distinctive. One major difference is that data used for predictive learning contains a designated class attribute, while descriptive data does not. Predictive tasks aim to learn the classification rules, whose heads contain the designed class attribute and collectively form a set of these rules as an induced model, based on a training dataset to classify a set of new records. Descriptive tasks aim to extract rules without a designed class attribute.

This review primarily focuses on investigating descriptive learning techniques, centralized around the APRIORI algorithm (Agrawal et al., 1993), because: 1) descriptive rule learning using APRIORI-like algorithms is very popular in spatial data mining; and 2) the adoption of descriptive individual rule learning to the predictive rule set learning is straightforward by considering only rules with the designed class attributes at the rule head. Predictive and descriptive rule learning, however, can differ to achieve effectiveness and efficiency, particularly for multi-class predictive problems. Fürnkranz and Kliegr (2015) comprehensively reviewed predictive rule learning, detailing the covering algorithm, its successors (e.g., AQ, PRISM, CN2, FOIL, RIPPER, PROGOL, etc.), a variety of search techniques (i.e., exhaustive versus heuristic) and search strategies (i.e., top-down, bottom-up, bidirectional).

The most popular descriptive rule learning approach is association rule learning (ARL) (Zhang & Zhang, 2002). The process of ARL can be decomposed into two sub-processes: 1) to determine the predicate sets whose occurrences exceed a predefined threshold in the database (i.e., frequent predicate sets) and 2) to generate rules from the frequent predicate sets. Since the second sub-problem is quite straightforward, the research has primarily focused on the first sub-problem. The solution to this problem involves generating candidate sets and checking for frequent sets. The APRIORI algorithm is the most popular one because of its efficiency during the candidate generation process with pruning techniques to avoid measuring unnecessary item sets, while guaranteeing completeness. The algorithm, however, requires multiple scans of the database. Many of its successors thus focus on advancing the phase of finding frequent predicate sets while overcoming the multiple scan requirement (see Zhang & Zhang, 2002 for a comprehensive overview).

3. The APRIORI Algorithm

The key idea of the algorithm is to generate all predicate sets with a minimum occurrence threshold using an exhaustive level-wise search. A support and confidence measure framework is used to decide if a set of predicates is frequent and a rule is significant or not. The support of rule $X \rightarrow Y$ is defined as the ratio of the number of data records that satisfies both X and Y to the total number of data records. The confidence of rule $X \rightarrow Y$ is defined as the ratio of the number of records that satisfy both X and Y to the number of records that satisfy only X. It is important that support is not confused with confidence. While confidence is a measure of the rule's strength, support corresponds to the statistical significance.

The algorithm generates all frequent predicate sets of size 1 (i.e., containing one predicate), then of size 2, size 3, etc., until frequent sets of size k are no longer found. Candidate generation is performed with multiple passes over the dataset. In a pass, the algorithm counts the candidate item sets by using only the item sets found frequently in the previous pass. The function to find the frequent set of the APRIORI algorithm is as follows:

Function FREQSET (UoL records)

P_k	Set of frequent predicate-sets of size k, where each member has two fields: (i) predicate-set and (ii) support count.
C_k	Set of candidate predicate-sets of size k (potentially frequent sets), where each member has two fields: (i) predicate-set and (ii) support count.

Input: UoL records with their predicates.

C_1 = set of all predicate-sets size 1; k = 1.

While (C_k) is not empty do //loop until no candidate predicate-set left

//remove all infrequent items from C_k (requires a database scan)

$P_k = C_k$ removing {all infrequent items}

//generate new candidates

C_{k+1} = {all size k + 1 predicate sets can be formed by uniting two predicate sets in P_k.}

$C_{k+1} = C_{k+1}$ removing {all sets for which not all subsets of size k are contained in P_k}

k = k+1

End of While

Answer = $\cup$ {P_k: k≥1}

4. Characteristics of Spatial Data for Rule Learning

Rule learning is a promising technique for mining patterns of correlations with spatial big data. Rule learning for spatial data will be referred to as spatial rule learning.

Spatial data includes spatial features that are georeferenced (i.e., their locations are determined within a geographic coordinate system). Thus, spatial data possesses spatial attributes embedded in feature locations on or near the Earth's surface. The nature of the geographic space, the complexity of the spatial object relationships, and the heterogeneous and sometimes ill-structured nature of geographic data, brings uniqueness to spatial rule learning. At the same time, it renders the standard rule learning techniques inefficient. Special characteristics of spatial information to be considered for spatial rule learning include:

1. appearances of object locations,
2. functional semantic and spatial relationships among objects (e.g., spatial effects and spatial interactions),
3. functional complexity posed by spatial dependency and spatial heterogeneity,
4. spatio-temporal changes of objects' semantic and spatial characteristics, and thus, their interactions, and
5. the heterogeneous and sometimes ill-structured nature of geographic data.

Without considering these spatial factors, classical rule learning approaches are a poor-fit to mining spatial data tasks (Koperski & Han, 1995; Mennis & Guo, 2009; Miller & Han, 2009).

Evolving from classical rule learning techniques, the objective of spatial rule learning is

then to extract the frequent occurrence of both semantic and spatial attributes of analyzed spatial features or of object locations (co-locations). Spatial predicates expressing spatial characteristics and relationships of the learning units are often used in addition to non-spatial predicates when applying the APRIORI algorithm. The process of materializing all possible spatial characteristics and relationships to generate a complete set of spatial predicates become crucial. This task is, however, non-trivial. The achievements and remaining challenges in spatial rule learning can be broadly discussed under two learning problems: 1) spatial association rules and 2) co-location rules.

5. Spatial Association Rule (SAR) Learning

SAR learning is a powerful technique that uses predicates to express the aspatial and spatial characteristics of analyzed features, as well as the spatial effects and spatial interactions among them,.

Formally, a SAR problem could be stated as: Let D be a spatial database containing spatial objects O. Let P be a set of all possible predicates, both non-spatial and spatial, that could be derived from D. Each object O has a unique identifier. Object O possesses a set of some predicates in P, representing the semantic or spatial properties of O. A spatial association rule is an implication of the form $X \rightarrow Y$, where $X \in P$, $Y \in P$, $X \cap Y \neq \emptyset$, and $\exists(x \in X \text{ or } y \in Y)|x,y$ are the spatial predicate. The rule $X \rightarrow Y$ holds in D with confidence c if c% of the objects in D that contain X also contain Y. The rule $X \rightarrow Y$ has support in D if s% of the objects in D contain both X and Y.

An example of using SAR learning is to mine associations to crime activity using block groups as the unit of learning. The data table to mine (Figure 1) contains rows being block groups with columns which contain block group IDs, non-spatial aggregated demographic variables (e.g., employment, crime incident counts, business store counts). Complex spatial attributes can be generated presenting the distance-based spillover effects of certain neighborhoods (e.g., neighborhood of high employment, around a shopping mall, etc). Continuous attribute values can be classified into nominal categories by discretization or fuzzy mapping techniques.

FID	Employment	Crime	Business	High Employment Spillover	Low Employment Spillover	Crime Spillover	Business Spillover	Mall Spillover
0	L	H	H	N	Y	Y	Y	Y
1	H	MH	L	Y	Y	Y	Y	Y
2	H	L	MH	Y	Y	Y	Y	Y
3	H	L	L	Y	N	N	N	N
4	L	MH	MH	Y	Y	Y	Y	Y
5	H	MH	MH	Y	Y	Y	Y	Y
6	M	MH	H	Y	Y	Y	Y	Y
7	M	H	H	Y	N	Y	Y	N
8	M	MH	L	Y	Y	Y	Y	Y
9	H	MH	L	Y	Y	Y	Y	Y

UoL | Neighborhood containment effects | Neighborhood proximity effects | Point proximity effects

Figure 1. Format of final singular relational table to mine SARs (Dao & Thrill, 2016).

By handling complex spatial predicates properly, multidimensional correlations reflecting

explicit and implicit spatial functional relations, as well as the compounding or mitigating effects of the semantic and spatial attributes, can be revealed through SAR learning (Dao & Thill, 2016).

SAR mining system representative achievements include GeoMiner (Koperski & Han, 1995), SPIN! (May & Savinov, 2003), SPADA (Lisi & Malerba, 2002), and SpatialARMED (Dao & Thill, 2016). SPIN! and SPADA, in particular, were developed to mine rules out of a multiple relational (i.e., many tables) database to deal with the issue of granularity with a multi-level concept hierarchical structure often encountered in spatial data.

One major challenge for SAR mining is to materialize all possible spatial relationships and present them using the form of predicates with meaningful linguistic values to facilitate the learning process. Many SAR learning systems (e.g., SPIN!, SPADA) have handled these challenges with an efficient indexing mechanism to extract non-complex spatial relationships (i.e., geometrical, topological, directional), while assuming a readily available complex concept hierarchy and relationships. SpatialARMED, in particular, facilitates the construction of complex spatial predicates posed by spatial effects and interactions (Figure 1) within and among neighborhoods. In this case, neighborhoods are defined as concentrations of high or low or homogeneous values regarding a certain attribute and detected using data-driven clustering techniques (see Dao & Thill, 2016; Dao & Thill, 2017 for details).

Fuzziness in the spatial concept hierarchy or relationship (e.g., near or far, large or small, strong or weak influence, high or low concentration) is another issue for SAR mining. It needs careful predication to derive meaningful predicate values and increase the chances of finding interesting rules (Laube et al., 2008). Efforts have also been made in handling spatial heterogeneity for SAR mining, including using subgroups and regional SAR mining. Here, the rule search space is divided into groups or regions based on a moving window in a similar manner to weighted regression (Li, 2008) or cluster analysis (Ding et al., 2011).

6. Co-location Rule Learning

Rather than being interested in associations among feature attributes and their functional relationships, co-location rule learning concerns the locations of analyzed features and aims to extract sets of spatial features often located near each other. In other words, co-location rules infer the co-presence of spatial features in the neighborhood of other spatial features.

The co-location rule learning problem is formalized as: Given 1) a set T of k spatial feature types T = {f1, f2, ..., fk} and their instances I = {i1, i2, ..., iN}, where each i is a vector containing instance-id, spatial feature type, and location, where location belongs to a spatial framework S and 2) a neighbor relation R over instances in I, efficiently find all the co-located spatial feature subsets c of T.

Nile Crocodiles à Egyptian Plover is an example of co-location rules. This rule states that the Egyptian Plovers (birds) are frequently found near Nile Crocodiles. As shown in Figure 2, data used for co-location rule learning are tables of found instances for each feature types X, Y, and Z represented by instance ID Boolean predicates.

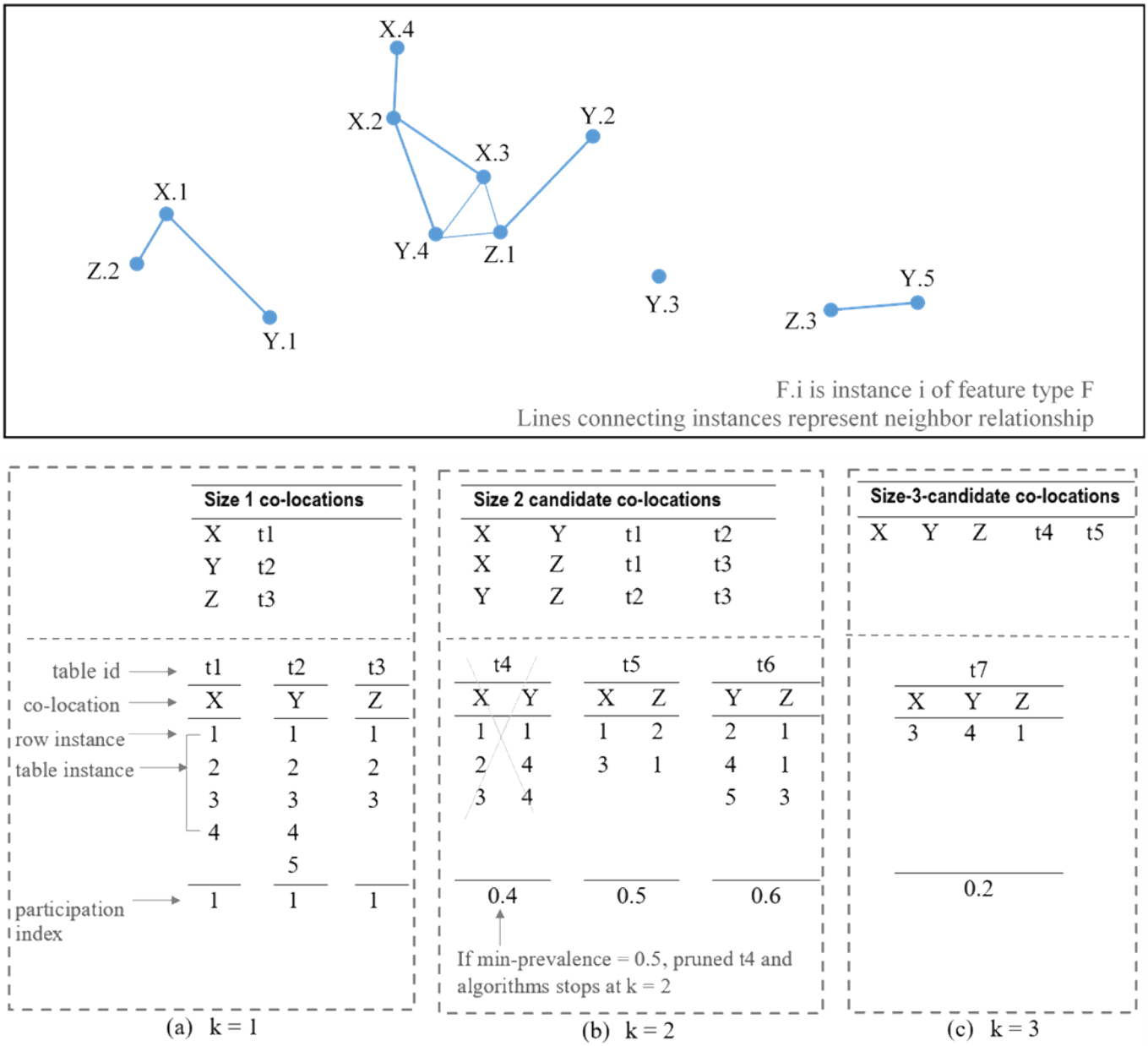

Figure 2. Spatial dataset to illustrate the event centric model for co-location rule learning.

Co-location rule learning utilizes the concept of neighborhood and APRIORI-like algorithms to mine sets of spatial features located near each other. A difference from spatial ARL is that there is a concept of co-location, instead of frequency. The support and confidence framework can no longer be applied. Instead, a neighborhood definition, along with the so-called prevalence measure (i.e., participation index), are used to decide whether the set of spatial features are co-located or not. In particular, when generating a set of candidate co-locations size k + 1 from co-locations size k, only the combination of features whose instances have neighborhood relationships are considered. After generating all candidate co-location sets of size k + 1, their prevalence measures are calculated and used to seek significant co-location. The participation ratio pr(c,fi) for feature type fi in a size-k co-location c = {f1, …, fk} is the fraction of instances of feature fi neighborhood-reachable to instances of other feature types in c. The participation index of the co-location c is the minimum participant ratio of feature types involved in c.

The earliest achievements in co-location rule learning relate to the co-location miner algorithm and its advancement, proposed by the Spatial Computing Research Group at the University of Minnesota led by Shashi Shekhar (Shekhar & Huang, 2001; Huang et al., 2004; Yoo & Shekhar, 2006). Later advancements for the co-location rule learning further refine the concepts and techniques to materialize neighborhood relationships of spatial features (e.g., by adopting k-nearest neighbor graphs (kNNGs) instead of distance thresholds (Qian, 2014)), and by defining network-based neighborhoods (Yu, 2016), to apply parallel processing frameworks for performance speedup (Yoo et al., 2016), and to mine rare co-location rules (Huang, 2006).

7. Concluding Remarks

Rule learning is a promising approach for spatial data mining to extract co-location patterns and correlations among the attributes of the analyzed features. Although the adaptation of traditional mining techniques to spatial data is not trivial, research efforts have demonstrated significant achievements in both spatial ARL and co-location rule learning. Future research in spatial rule learning could focus on the use of domain knowledge to develop effective rule evaluation, context-based learning techniques, blended or hybrid knowledge learning techniques, as well as the development of spatio-temporal rule learning techniques. Regarding the latest, readers are referred to further readings on sequential pattern learning and sequential rule learning (e.g., Agrawa & Srikant, 1995; Huang et al., 2008; Fournier-Viger et al., 2017). Sequential rule learning is particularly promising for various spatio-temporal applications such as those relating to changes, movements, and interactions. Among many others, the APRIORI algorithm can be coupled with temporal information to extract sequential frequent sets in order to generate sequential rules (e.g. Fournier-Viger et al., 2012).

References

Agrawal, R., & Srikant, R. (1995). Mining Sequential Patterns. Paper presented at the Proceedings of the Eleventh International Conference on Data Engineering, Taipei, Taiwan.

Agrawal, R., Imieliski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. Paper presented at the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., United States.

Dao T.H.D. & Thill, J-C. (2017). Detecting Attribute-Based Homogeneous Patches Using Spatial Clustering: A Comparison Test. In Popovich, V., Schrenk, M., Thill, J. C., Claramunt, C., Wang, T. (Eds.) Information Fusion and Intelligent Geographic Information Systems (IF&IGIS'17). Lecture Notes in Geoinformation and Cartography. Springer, Cham.

Dao, T. H. D., & Thill, J-C. (2016). The SpatialARMED Framework: Handling Complex Spatial Components in Spatial Association Rule Mining. Geographical Analysis, 48(3), 248-274.

Ding, W., Eick, C. F., Yuan, X., Wang, J., & Nicot, J.-P. (2011). A Framework for Regional Association Rule Mining and Scoping in Spatial Datasets. GeoInformatica, 15(1), 1-28.

Fournier-Viger, P., Faghihi, U., Nkambou, R., & Nguifo, E. M. (2012). CMRules: Mining sequential rules common to several sequences. Knowledge-Based Systems, 25(1), 63-76.

Fournier-Viger, P., Lin, J. C.-W., Kiran, R. U., Koh, Y. S., & Thomas, R. (2017). A Survey of Sequential Pattern Mining. Data Science and Pattern Recognition (DSPR), 1(1), 54-77.

Fürnkranz, J., Gamberger, D., & Lavrač, N. (2014). Foundations of Rule Learning: Springer.

Huang, Y., Pei, J., & Xiong, H. (2006). Mining Co-Location Patterns with Rare Events from Spatial Data Sets. GeoInformatica, 10(3), 239-260.

Huang, Y., Shekhar, S., & Xiong, H. (2004). Discovering colocation patterns from spatial data sets: a general approach. In IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 12, pp. 1472-1485.

Huang, Y., Zhang, L., & Zhang, P. (2008). A Framework for Mining Sequential Patterns from Spatio-Temporal Event Data Sets. IEEE Transactions on Knowledge and Data Engineering, 20(4), 433-448.

Koperski, K., & Han, J. (1995). Discovery of spatial association rules in geographic information databases. In: Egenhofer, M.J., Herring, J.R. (eds) Advances in Spatial Databases. SSD 1995. Lecture Notes in Computer Science, vol 951. Springer, Berlin, Heidelberg.

Laube, P., Berg, M., & van Kreveld, M. (2008). Spatial Support and Spatial Confidence for Spatial Association Rules In A. Ruas & C. Gold (Eds.), Headway in Spatial Data Handling (pp. 575-593). Berlin Heidelberg: Springer

Li, X. (2008). Mining Spatial Association Rules in Spatially Heterogeneous Environment. Paper presented at the International Conference on Earth Observation Data Processing and Analysis (ICEODPA), Wuhan, China.

Lisi, F. A., & Malerba, D. (2002). SPADA: A Spatial Association Discovery System. Paper presented at the 2002 International Conference on Data Mining, Bologna, Italy.

May, M., & Savinov, A. (2003). SPIN!–an Enterprise Architecture for Spatial Data Mining. In: Kovalerchuk, B., Schwing, J. (eds) Visual and Spatial Analysis. Springer, Dordrecht.

Mennis, J., & Guo, D. (2009). Spatial Data Mining and Geographic Knowledge Discovery - An Introduction. Computers, Environment and Urban Systems, 33(6), 403-408.

Miller, H. J., & Han, J. (Eds.). (2009). Geographic Data Mining and Knowledge Discovery: An Overview. CRC Press, Taylor and Francis Group.

Qian, F., Chiew, K., He, Q., & Huang, H. (2014). Mining Regional Co-location Patterns with

kNNG. Journal of Intelligent Information Systems, 42(3), 485-505.

Shekhar, S. & Huang, Y. (2001). Discovering Spatial Co-location Patterns: A Summary of Results. In C. S. Jensen, M. Schneider, B. Seeger & V. J. Tsotras (Eds.), Advances in Spatial and Temporal Databases: Proceedings of the 7th International Symposium, SSTD 2001 Redondo Beach, CA, USA, July 12-15, 2001 (pp. 236-256). Berlin, Heidelberg: Springer.

Yoo, J. S., & Shekhar, S. (2006). A Joinless Approach for Mining Spatial Colocation Patterns. IEEE Transactions on Knowledge and Data Engineering, 18(10), 1323-1337.

Yoo, J. S., Boulware, D., & Kimmey, D. (2014). A Parallel Spatial Co-location Mining Algorithm Based on MapReduce. 2014 IEEE International Congress on Big Data, Anchorage, AK, USA, 2014, pp. 25-31.

Yu, W. (2016). Spatial Co-location Pattern Mining for Location-based Services in Road Networks. Expert Systems with Applications, 46, 324-335.

Zhang, C., & Zhang, S. (2002). Association Rule Mining: Models and Algorithms. Springer-Verlag.