MINING FUZZY SPATIAL CONFIGURATION RULES: METHODS AND APPLICATIONS

Rongqin Lan [a, *], Wenzhong Shi [b], Xiaomei Yang [c], Guangyuan Lin [d]

[a] Zhengzhou Institute of Surveying and Mapping, Zhengzhou City, Postal Code: 450052 – Lanrq@sohu.com
[b] Land Surveying and Geoinformatic Department, The Hong Kong Polytechnic University – Wenzhong Shi@polyu.edu.hk
[c] Urban Construction Information Centre of Xiamen City, Xiaman, Postal Code: 361004
[d] The State Key Laboratory of Resources and Environmental Information System (LREIS), Beijing, Postal Code: 100101－Yangxm@lreis.ac.cn

KEY WORDS: Spatial Configuration Pattern, Spatial Association Rule Mining, Spatial Analysis, Fuzzy Spatial Relation Reasoning

ABSTRACT:

In view of the related research results of GIS and spatial analysis and according to the requirement of fuzzy spatial queries, this paper presents a spatial configuration rule mining and feature extracting method based on spatial association rule mining and fuzzy spatial relation reasoning. We detailed introduce the principle and method as well as application examples. A six mining steps established as follow: acquiring spatial data; selection of spatial predicates and knowledge representation; object recognition and extracting based on machine learning; reasoning objects' spatial relation; extracting spatial frequent itemset; discovering frequent patterns; synthetically evaluate the mining result in contrast to the query. Examples and their mining results were presented to illustrate the algorithm's validity.

1. INTRODUCTION

Spatial configuration pattern refers to the arrangement law, distribution state or construction way of spatial objects' in space. A spatial configuration rule is the recognition and presentation of this kind of rules. It is very useful for spatial analysis, geographic information system (GIS), remote sensing image classification and many others spatial application areas which needed to modelling objects' spatial distribution pattern. For example, when we query through GIS with "find all patterns which object 1st locates in about 10km southeast object 2nd and cross with object 3rd"[4],[5].

Spatial configuration pattern mining is to find the geospatial distribution law by using the methods and technologies of spatial data mining. The main task is to distinguish and extract some hidden, unknown and interesting spatial/temporal distribution patterns in real world. The types of knowledge it can discover are as follows: the regulations of geospatial distribution patterns in certain area; spatial objects or events that are interested and repeated; constructing a spatial distribution that can gratify the need of user's spatial query; predicting the trend of spatial objects or events.

There are three distinguished features between spatial configuration pattern mining and other methods (for examples, pattern recognition and mathematical morphology). Firstly, spatial data mining is data-driven. We have no the knowledge of spatial pattern beforehand. Secondly, data are in scattered way, it is not capable of Sets operation for image. Finally, when a new spatial arrangement pattern found, its spatial feature must be described for spatial reasoning.

There are many ways to find spatial configuration rules. We can integrate machine learning with spatial frequent items mining to discover a spatial configuration pattern hidden in vector GIS data. Other ways include spatial cluster mining, spatial statistics, spatial auto-regression analysis, etc[2].

2. SPATIAL CONFIGURATION PATTERN MINING BASED ON SPATIAL FREQUENT ITEMSETS

Generally, there is a causal relationship between the spatial position of object and its surrounding environment. This kind of causal relationship can be discovered by spatial association rule mining. The discovered knowledge is very useful for widely area.

Spatial association rule describes the intrinsic law of co-occurrence of spatial objects. In other words, it describes the conditional frequency of co-occurrence of data items of spatial entities in a given spatial database. It mainly refers to the association rule of neighbouring, conjoining, accreting and containing among spatial entities[7],[11].

Essentially, it can be defined as following:

$$X \Rightarrow Y\ (s\%, c\%, D) \qquad (1)$$

Where, s% and c% respectively stands for support and confidence of the rule. D is the distance threshold between two spatial objects. If two spatial objects X, Y satisfy the distance (X, Y) < D, then X, Y satisfy the generalized spatial neighbourhood relation.

Spatial association rule mining is the extension of transaction association rule. Unlike traditional association rule mining, there exists extra metric information in the definition of spatial association. Accordingly spatial association rule mining is divided into two steps, first of which is to find all frequent itemsets of candidates in a neighbouring field with a metric

* Corresponding author. Lan Rongqin, Longhai Middle Road 66#, Zhengzhou City, 450052, Tel. 0371-63535055.

threshold, while the second is to generate rules from frequent itemsets with min-support and min-confidence. It is apparently that the metric is the most important measure for the spatial objects' association. According to the definition of spatial association rule, the so-called 'co-occurrence', in comparison to time sequence, refers to the space proximity of objects. When mining a transaction database, zero value is assumed for the distance of goods in customs' basket. But for spatial objects, the relationship between roost and wellspring must be measured with spatial interval or orientation. For this reason, the most important characteristic of spatial data mining is the use of spatial predicates, such as topological predicate, metric predicate and directional predicate. As an important index of association, the distance metric is usually discrete and replaced by several spatial relation predicates, for examples, far, near, meet, within_distance_of, beyond_distance_of, across, overlap, etc.

K. Koperski (1999) generally introduced the principle and application on spatial association rule mining [7]. K. Koperski, J. Han (1995) proposed transaction-based approaches [11]. Yan Huang et all (2004) proposed distance-based spatial association rule mining and its application [14], and simultaneously developed a algorithm for mining co-location pattern from large spatial database. Esen Kacar (2001) proposed fuzzy topological relation mining [13]. Isabelle Bloch proposed a new definition of the relative position between two objects in a fuzzy set framework [15]. George Brannon Smith et all (2002) processed spatial relation hierarchy and concept hierarchy with fuzzy membership and proposed fuzzy spatial association rule mining method [12]. These investigations and determinations indicate that spatial association rule mining has formed a perfect theory framework and had a variety of flexible algorithms. In advance, it can be used to discover hidden geo-distribution patterns, correlation characteristics or interaction rules.

It is apparently that spatial association rule mining can be used to discover the spatial or non-spatial association relationship among spatial entities on map. From frequent itemsets with causal relationship we can test and extract spatial configuration pattern and discover some meaningful and potential arrangement pattern. Spatial configuration pattern mining based on spatial association rule mining can be treated as spatial relationship-constrained spatial association rule mining. It generally includes the following steps: extracting spatial object, machine learning, reasoning the relationship of spatial objects, extracting the frequent intemsets, discovering the pattern, similarity measuring.

2.1 Acquiring spatial data

The test data are generally from existing GIS data. Here we use several large scale topographic maps in certain area. The land classifications mainly include river, road, cropland and village.

2.2 Predicate Selection and Knowledge Presentation

Spatial or non-spatial relationship should be presented with spatial predicates. In this paper, the selected predicates are as figure 1.

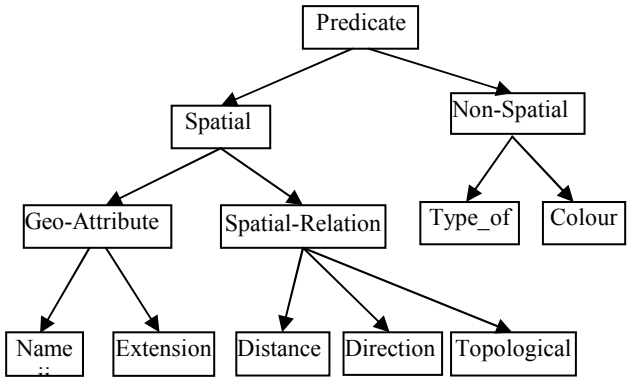

Figure 1. Hierarchy of spatial predicates

Spatial entities and their relationships are presented with attribute, predicate and function. For examples,

$$adjacent\ (X, Y)$$
$$extension\ (Z) \in [182.9..98.1]$$
$$f_p(X, Y) = true$$

2.3 Spatial Object Recognition and Extracting based on Machine Learning

GIS describes the geographic objects by means of the physical entity, e.g. point, line or polygon[1]. Although signs or symbols in GIS correspond to general concepts (for example, river, boundary, downtown). But some geo-referenced concepts might not have been explicitly modelled by map creator in database. For instance, a map's qualitative interpretation may require the recognition of steep and gentle slopes, these data are not stored in the repository and it may be hard to formulate a query that return them. For the purpose of object recognition and extracting, we are interested in machine learning that satisfy our requirements.

Machine learning is often referred to as fundamental tools of data mining or knowledge discovery, which is a process of finding useful information in large quantities of raw data. The procedure of training machines to do automatic knowledge discovery can be conceived as a search through the large space of examples. Each example is typically represented by a vector of attribute-value pairs, either numeric or symbolic. The individual examples match with various concepts that are either known or unknown before learning. If the concepts are known, then supervised machine learning algorithms are applied to extract a maximally general description that should cover all examples of a concept and exclude all other examples. In general, two essential search operators, generalization and specialization, are iteratively used. Generalization tries to make a hypothesis for a concept more general whenever a new example of the concept arrives, whereas an example that does not match the concept can necessitate the specialization operator. In case of unknown concepts, unsupervised machine learning algorithms are applied to find conceptual clusters in examples according to certain optimization criteria and then prompt users to assign a concept name for each cluster.

In our study, we try to train a machine through examples and background knowledge in a way analog to human learning. The machine learning process is as Figure 2.

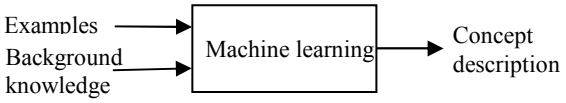

Figure.2 Machine learning process (Kubat, et al 1998)[6]

First of all, we should develop a set of qualitative reasoning rules to act together with background knowledge. The system then will derive all kinds of physical objects and relations that did not explicitly represented in the logical description of observations. The rules take form of $L_1, L_2, \cdots L_m \rightarrow L_0$, where the conjunctive premise part $L_1, L_2, \cdots L_m$ is called the body, while the conclusion part L_0 is called the head. Secondly, a set of observations O are described in a language $\mathfrak{L}_0$. For instance,

Colour predicate $colour(X) = [blue, red, black ...]$
Extension predicate $extension(X) \in [Num01, Num02]$
Symbol type predicate $type\_of(X) = [po\text{int}, line, ...]$

Finally, learning system will reason a set of concepts $C_1, C_2, \cdots C_r$ to be learned from topographic map of certain mountain area based on background knowledge and description languages. It then recognizes the type of physical objects as the form of:

IF $colour(X) = blue$
AND $extension(X) \in [203.17, 428.49]$
AND $type\_of(X) = line$
THEN $class(X) = river$

2.4 Reasoning the relation of geographic objects

There are three relations between point, line and polygon: distance metric relations, topological relations and directional relations. The judgement models of distance metric relations mainly include: distance between point to point, distance between point to line, distance between point to polygon, distance between line to line, distance between line to polygon, distance between polygon to polygon, etc. The judgement models of topological relations include: X through Y, X cross Y, X disjoint from Y, X covers Y, X is inside Y, X equals Y, etc. The judgement models of directional relations include: point to point/line/polygon, line to line/polygon, polygon to polygon, etc. Generally, distance metric relations modelling for spatial data mining use the measure "mean distance". So the calculation is relative simple. Here we derive the calculation equation of directional relations.

Without loss general, assume the position of a geographic object be located in centroid. The direction between two geographic objects is the angle of two centroids. In figure.3, we divide space into 8 orientations.

$$A = \{East, NorthEast, North, NorthWest, West, SouthWest, South, SourhEast\} \quad (2)$$

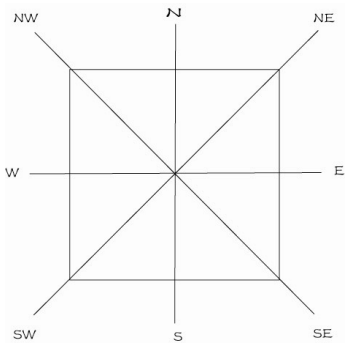

Figure 3. Definition of direction relation

while

$A_1 = NorthEast\ (22.5^0, 67.5^0)$
$A_2 = North\ (67.5^0, 112.5^0)$
$A_3 = NorthstWes\ t(112.5^0, 157.5^0)$
$A_4 = West\ (157.5^0, 202.5^0)$
$A_5 = SouthWest\ (202.5^0, 247.5^0)$
$A_6 = South\ (247.5^0, 292.5^0)$
$A_7 = SouthEast\ (292.5^0, 337.5^0)$
$A_8 = East\ (332.5^0, 22.5^0)$

If the angle of two centroids of geographic objects is θ, directional relation is A_i, and then we can get the similarity of direction:

$$\sigma_A = \begin{cases} \frac{\theta - 22.5^0 * j}{22.5^0} & 22.5^0 * j < \theta < 22.5^0 * (j+1) \\ \frac{22.5^0 * (j+2) - \theta}{22.5^0} & 22.5^0 * (j+1) < \theta < 22.5^0 * (j+2) \\ 1 & \theta = 22.5^0 * (j+1) \end{cases} \quad (3)$$

where $j = 2 * i - 1$, j is odd number.

We can calculate the direction similarity of arbitrary two geographical objects according to equation (3). For instance, if the angle of two centroids is 230^0,

since j= (230/22.5-1) =9.2≌9, so i=5

i.e., one object is to the southwest of another object.

Veracity of the assertion is 78%, that is,

$$\sigma_A(A_5, 230^0) = \frac{22.5^0 * 11 - 230^0}{22.5^0} = 0.78$$

Common used spatial topological predicates can be found in literature [3], [7]. In the case of fuzzy spatial data mining, fuzzy logic was introduced to provide a way of modeling the uncertainty of natural language. We construct a fuzzy spatial relation hierarchy by defining a partial belonging of a sub-item (predicate or spatial operator) to its parent items. For instance, the returned feature of spatial operator "Contains", some can be perceived as "Overlap", others can be perceived as "Cover". So *Contains* operator can be defined fuzzily as the child of both "Overlap" and "Covers" in the spatial relation hierarchy. In the same way, we can fuzzily define "Inside" as the child of both "Overlap" and "Covers". Hence, it is true that a predicate belongs to its parent predicates with different weights.

By means of above spatial relations (distance, topologic and directional), we can represent spatial knowledge and derive new, unknown, interesting spatial relations with spatial predicates. For examples,

$$close\_to(X,Y) = true$$
$$relation(X,Y) = almost\_parallel$$
$$geographic\_direction(X) = north\_east$$
$$distance(X,Y) = 474.19$$

2.5 Extracting spatial frequent itemsets

Before conducting mining, we must classify the spatial database and export geographical feature space $F, F = \{f_1, f_2, \cdots f_m\}$, which fi={$ei1$,$ei2$,…eii,…eim}; each entity $e_k \in E$ is a triple tuple <*ID*, *Feature Type*, *Location*>.

The first step is, according to distance threshold D, to generate 2-size pair of entities and construct mining database from discrete objects stored in original spatial database.

The famous Apriori algorithm is a capable mining association rule. First it scans original records and takes them as candidate c_1 to find 1-size itemset L_1 with min-support. Then the L_1 is used to generate candidate C_2 by computing $L_1 * L_1$. And then L_2 from C_2 with min-support can be obtained. That is,

$$L_2 = L_1 * L_1 = \{f_1 \cup f_1 \mid f_1, f_2 \in L_1 \mid f_1 \cup f_1 \mid = \phi\} \quad (4)$$

And K-size candidates can be calculated from the following equation,

$$L_k = L_{k-1} * L_{k-1} = \{f_i \cup f_j \mid f_i, f_j \in L_{k-1} \mid f_i \cup f_j = k-1\} \ (i \neq j) \quad (5)$$

The rest may be deduced by analogy.

The support S_p and confidence C_f of rule $X \Rightarrow Y$ are respectively as:

$$S_p = \frac{\sigma_{x \to y}}{|D_m|} \bullet 100\% \quad (6)$$

$$C_f = \frac{\sigma_y}{\sigma_x} \bullet 100\% \quad (7)$$

Where $\sigma_{x \to y}$, σ_x, σ_y respectively denote the support number of itemset $X \to Y$, X, Y, D_m is all transaction numbers of mined database.

2.6 Extracting patterns

After mining rules at the highest level of the spatial relation hierarchy, we can get the large k-predicate frequent itemset. Then the pattern of the large itemsets is just the mapping of arrangement of geographic objects in real world when we observing. To gain this ends, several steps should be taken,

1. Constructing image schema. Imaging the configuration pattern of geographic objects according to spatial query.
2. Defining kernel components. Searching space by taking these components as centre, and then searching further more by other given conditions of spatial relation.
3. Checking all satisfied patterns and labelling the graphs.
4. Determining the spatial positions and attributes of geographic objects, and storing them in spatial database.

3. FUZZY COMPREHENSIVE JUDGEMENT OF MINING RESULTS

One of fundamental functions of GIS is spatial query. In practice we may encounter the following query languages:

Example 1. Find the spatial relation between a certain road and a reservoir.

Example 2. Find a cliff which is adjacent to sea.

Example 3. Find all configurations where village A is *about 5km northeast* of valley B, which, in turn, is *inside* valley B.

Example 4. Find the situations in which all objects in one class tend to be located in the same direction relative to all objects in a second class.

There is no doubt that all these queries may return a large set of results from GIS. But if constraints (for *example* 1, a road is to the northwest of the reservoir) be added, then a solely result can be gotten.

Due to the diversity of solutions, it is necessary to apply spatial relation constraints. To analyze and compare the similarity between mining result and real spatial configuration, the fuzzy membership might be given according to the measurements of above three spatial relations. And then conduct fuzzy comprehensive judgment of mining results according to correlative fuzzy theory.

4. APPLICATION EXAMPLE

In the early stage of building a reservoir, to find an appropriate site is an important work. Since field exploration is expensive and time-consuming, it is very useful to make use of the potential of GIS and apply the function of spatial analyse. If we can find the underlying configuration pattern fitting for building a reservoir based on the type of land, ruggedness, gradient, then we can adjust measures to local conditions to locate the best site of dam.

To demonstrate the feasibility of spatial configuration pattern mining, several large scale topographic maps of certain mountain area are collected, as figure 4. Steps and methods are as follows.

Figure 4. Topographic map

Firstly, determine underlying and correlative entities (land type). Generally, ideal reservoir site is in the region of depression or intermontane basin which will make the body of water as large as possible. That is said, the configuration pattern will take the shape of “pocket”, “ampullaceous” or “oval”. While the correlative entities are mainly surrounding rivers. According to section 2.3, these objects will be picked.

Secondly, determine the association of selecting spatial entities. The statistic of contours and rivers are placed in table 1. Then according to section 2.4, spatial predicates will be used to

represent and reason the spatial relations. For instance, operator "across" was used to represent the cross between rivers and contours, and "close_to" used to express the neighbourhood, etc.

Object	Reservoir	Pocket	Oval	G2 river	G3 river	Other
Number	25	25	18	25	127	9

Table 1. Statistic of selecting objects

After calculating frequent itemsets according to equation (4), (5), (6), (7) in section 2.5, the results show that the most possible configuration pattern is composed of pocket contours, oval contours and grade-2 rivers. And its support and confidence is respectively 96.5% and 90.3%.

Thirdly, determine the configuration pattern of spatial associated entities. The pattern must be a formulated graph. That is, it is an abstract representative, not always the same as ground true. As seen in figure 5.

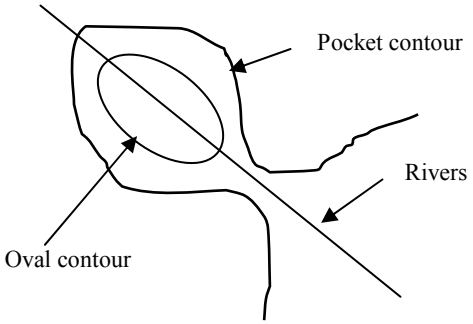

Figure 5. Formulated configuration pattern like "goldfish"

Finally, it is necessary to synthetically evaluate the mining result in contrast to the query according to distance, direction and topologic relations, and give the evaluating result. Here the process is omitted.

5. CONCLUSION AND DISCUSSION

In the research process, the difficulty of mining spatial configuration can be concluded as:

1. Finding frequent items in vector spatial data can be a great expense of time and effort. High efficient algorithms are needed.
2. Except the frequent patterns, some specific pattern is very useful. But they are likely to be omitted in our procedure.
3. Evaluating the similarity of mining result in contrast to the query is a very important step. But, as we know, since spatial relation is very complicated and rigor modeling spatial pattern is difficult, this work must be further discussed.

6. REFERENCE

[1] Wang jiayao. *Principles of Spatial Information System*. Beijing: Science and Technology Press.2001. pp236~239.

[2] Du Yunyan. *Geographic Case Reasoning an Its Applications*. A thesis submitted in partial fulfillment of the requirement for the degree of doctor in Chinese Academy of Science and Technology. 2001. pp13~25.

[3] Harvey J.Miller and Jiawei Han. *Geographic Data Mining and Knowledge Discovery*. London：Taylor & Francis, 2001. pp291~314

[4] S. Shekhar, P. Schrater, W. Raju, and W. Wu, "Spatial Contextual Classification and Prediction Models for Mining Geospatial Data", *IEEE Transactions on Multimedia*, 4(2): 174-188, 2002.

[5] Josef Sivic, Andrew Zisserman. Video Data Mining Using Configurations of Viewpoint Invariant Regions. Http://www.robots.ox.ac.hk/~vgg.

[6] Kubat, M. et al, 1998 "A review of machine learning methods" in "Machine Learning and Data Mining" Methods and Applications, Ed. Michalski R.-S. et al, John Wiley & Sons, pp.3-69.

[7]Krzysztof Koperski. A progressive refinement approach to spatiall data mining. A thesis submitted in partial fulfillment of the requirement for the degree of doctor of philosophy in the school of computing science, Simon Fraser University. April, 1999.

[8] Guoqing Chen, Qiang Wei, Etienne Kerre, Fuzzy Data Mining: Discovery of Fuzzy Generalized Association Rules[A]. In Bordagna & Pasi (eds.), *Recent Research Issues on Management of Fuzziness in Databases*. Physica-Verlag (Springer), 2000.

[9] G. Chen, Q. Wei, E.E. Kerre. Fuzzy Data Mining: Discovery of Fuzzy Generalized Association Rules. *Fuzzy Terms*, pp. 45-66.

[10] Guoqing Chen, Geert Wets, Koen Vanhoof. Representation and Discovery of Fuzzy Association Rules (FARs). www.luc.ac.be/iteo/Representation.doc.

[11] K. Koperski, J. Han. Discovery of Spatial Association Rules in Geographic Information Databases. *Symposium on Spatial Databases*. 1995.

[12] George Brannon Smith, Susan M. Bridges. Fuzzy Spatial Data Mining. www.cs.msstate.edu/~bridges/papers/nafips2002b.pdf

[13] Esen Kacar, Nihan Kesim Cicekli. Discovering Fuzzy Spatial Assciation Rules. www.ceng. metu.edu.tr/~nihan/pub/esen.ps.

[14] Yan Huang et all. Discovering Colocation Patterns from Sparial Data Sets: A General Approach. *IEEE Transactions on Knowledge and Data Engineering*. Vol.16, No.12,December 2004.

[15] Isabelle Bloch. 1999. Fuzzy relative position between objects in image processing: A morphological approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no.7, pp. 657-664.