Collocation Pattern Analysis: A Variable Size/Shape Analysis

M. Nagabhushana Rao +

*+Professor **of Computer Science,***
C.R..Engineering College,
Tirupati, India.

Dr. P. Govindarajulu ++

++Professor of Computer Science,
Sree Venkateswara University,
Tirupati, India.

Abstract:

Spatial Data Mining is discovering of interesting, implicit knowledge in spatial databases, an important task for understanding and use of spatial-data and knowledge bases. The spatial features are identified in the spatial data and various kinds of analyses are aimed at obtaining various results by the decision support systems. Collocation patterns represent subsets of fuzzy/boolean spatial features whose instances are often located in close geographic proximity. Several theoretical and application-oriented approaches support finding the collocation patterns. The analysis gives the scope of understanding the variable changes of spatial features that evolve in the collocation patterns and how they influence the change in the spatial collocation rule.

Keywords: spatial data mining, collocation pattern, and spatial knowledge

1. INTRODUCTION

Spatial data mining becomes more interesting and important as more spatial data have been accumulated in spatial databases. Spatial patterns are of great importance in many GIS applications that yield equal to association rules of a business (OLTP). Mining the spatial collocation patterns is an important spatial data-mining task with broad applications. Certain operations that incorporate methods of analyses and summarization are very much crucial for organizations having large data sets of spatial data.

Association rule finding is an important data mining technique, which helps retailers finding items frequently, bought together to make store arrangements, plan catalogs, and promote products together. For decision support systems to get enriched with information like changes and trends occurred in the spatial zones. Particularly, the observations on the representation of collocation pattern and its varying size using semantically supported elements is of more important to archeologists, GIS scientists, governments for analyzing the changing trends in the civilization. Many spatial datasets consist of instances of a collection of Boolean spatial features. Spatial association statistics measures the concentration of an attribute over a space.

The task of observing the changes in the co-location pattern incorporates several approaches and principles. Each method or a procedure of its kind can evolve in identifying or obtaining the inferences from the co-location patterns.

2. RELATED STUDY

According to the immense definitions made in [1];
Definitions:

1. *A collocation is a subset of Boolean spatial features or spatial events.*
2. *A spatial association rule derivative, the collocation rule is of the form:*

$$C_1 \rightarrow C_2 (p, cp)$$

C1 and C2 are collocations, p is a number representing prevalence measure and cp is a number measuring conditional probability [1,3,6,7].

The prevalence measure and the conditional probability measure are called interest measures and are defined differently in different models. Research on mining spatial associations or collocations is based on two models; *the reference feature model* and *the collocation pattern model*; however they have their pros and cons. In this section we refer to the co-location pattern model and describe the collocation derivation and then suggest a probabilistic approach to study the changes.

2.1 COLLOCATION PATTERN MODEL

In the fundamental phase (preprocessing) the patterns often called as spatial patterns that are mined from the raw spatial data, which contains lot of hidden spatial information. The interesting patterns that describe the generality of rule in spatial data among the mined spatial patterns are defined as collocations or the collocation patterns. The research has ushered towards mining collocation patterns, which are feature sets with instances that are located in the same neighborhood [1,3,4,6,7]. The following equation describes the participation ratio *pr* (f_i, **P**) of a feature f_i in the pattern **P**.

$$pr(f_i, P) = \frac{\#[f_i][P]}{\#[f_i]}$$

Using this measure, we can define collocation rules that associate features with the existences of other features in their neighborhood.

3. ANALYSIS

Spatial collocation rule mining is sometimes presented as a one-off exercise [2]. This is a misconception. Rather, it should be perceived as an on-going process (even if the spatial feature set is fixed). One examines the collocation rule one way, interprets the results, looks more closely at the collocation rule from a related perspective, looks at them another way, and so on. The point is that, except in those very rare situations when one knows what sort of changes takes place in collocation rule is of interest, the essence of collocation rule mining is an attempt to discover the unexpected - and the unexpected/unexplored, by its very nature, can arise in unexpected ways.

3.1 MINING SPATIAL DATA DEVIATION AND EVOLUTION RULES:

The analysis in this paper is carried out about the trends of a spatial zone that tends to change the well-formed spatial collocation rules in a evolutionary way. Evidential reasoning should be explored in the mining to find the uncertainty in the evolution of the new spatial collocation rules or the evolutionary changes made to the spatial collocation rules.

The defining attribute of the spatial data mining is that the spatial data is very large, and the spatial domain is typically considered to be as superfluous geographic space. There is a strong possibility that collocation rule will undergo several changes in the spatial domain [1,3,6,7,9]. The changes are with respect to the addition or elimination of attributes (boolean/fuzzy spatial features) for a spatial object that forms a collocation pattern. Often, spatial domain changes its boundaries while the process of analyses, to enable the Geo-users use most of the space in the geography. More number of liable situations prevail that cause changes in the attributes, when domain specification of the spatial zone is varied. The geographic data stream has to be reviewed many times in order to identify the relevant changes take place in the collocation rules. The relationship between every change observed leads to the derivation of phenomena to the Geo-users. To make it more understandable a fuzzy set concept of the Boolean spatial features referring to the spatial object are considered to be apt [3].

The fuzzy set concept in this analysis seems inevitable to study. For each attribute, according to Boolean principles, have two states to identify the attribute as present or absent in the pattern. The fuzzy set boosts the clarity that how much of the attribute is influencing the pattern. Some attributes may coarsely influence and some may densely influence the pattern. The spectrum of this varied presence of the attribute can be considered to identify the different states of the collocation pattern.

However, the changes are identified as different states of a collocation rule. There is no basis for the states which are the changes embodied in the collocation and they are evolutionary to the changing trends and evolving features of the spatial domain. That is there is no particular regularity what state element of the fuzzy set can be attributed to the spatial feature. New combinations of new set of features give raise of new phenomena in the pattern. The phenomenal rule system may also be implemented for the definition of varying collocation size.

As there is a reason to analyze the changes in the collocation and they are very large in number, and they are used for wide variety of applications, the spatial data with many important changes made in the trends over time because of the changes in the underlying phenomena. The process is called *spatial data evolution*. Because of the changes made to the collocation, as the user may be able to glean valuable insights into emerging trends in the underlying spatial activity.

The accent of statistics on mathematical rigour can reveal heavy emphasis on the concept of *inference* [2]. A simple approach to the theory of data mining is to declare that data mining is statistics and thus search for theoretical framework for data mining can stop immediately: The theory of data mining is statistics. The statistical model built in this paper may make use of a sequence of probability statements.

A possible theoretical approach to spatial data mining is to view data mining as the task of finding the underlying joint distribution of the variables of the data. To give a probabilistic model for the phenomenon of spatial data mining, we have to define the number of instances of changes, identification of changes done to the patterns, for each instance. The interesting point to notice is a **Figure 1** could be interpreted as part of the probabilistic model.

Fig.1 Probabilistic Model

From the above diagram, we can infer that based on the demands of the spatial domain the *trends* occur inadvertently, as changes due to demands sway on, the *features* evolve, and in turn as they are the pack of *patterns*, the *collocations* change.

Let us consider **C** is the collocation, contains a set of related features, $f_1 \ldots f_k$. The collocation is referred to as $\mathbf{C}\{f_1 \ldots f_k\}$. For identity of the collocation, a collocation lead is assumed. The features of a collocation are of various degrees of importance. A high degree feature can be represented as collocation lead. Based on the intensity and weights of features a numerical value is assumed. The collocation lead will be a heavy weight number, called as collocation lead number. The set or series of collocation lead numbers are the fix to a random variable, which again is a result of the random variable.

4. PROBABILISTIC DEFINITION

The random phenomenon of changing features in a geographical space is the random experiment representing the probabilistic framework that is followed here. The totality of changes is the possible outcomes of the experiments that denote entropy, are the elements of the sample

space. Consider, the elements of sample space are the various elements in the spatial domain of one or more dimensions, which are the outcomes, generated from experiments.

The infinite sample space for the experiment is considered as the entropy of spatial domain, which is represented by $\mathcal{E}$. Patterns contain feature-sets that belong to the elements of space. The conditional probability of patterns in spatial domain can be defined as

$$S = P\{\mathcal{E} \mid A(\mathcal{E})\} \rightarrow (1)$$

S is the resultant that contains the probability conforming possible spatial objects in patterns. Consistently, pattern is a collocation of spatial objects. More importantly the pattern highlights the features of spatial objects. The probability spatial domain that contains collocation, which complies with feature set called A, is given in the equation (1).The features, the basic and essential properties tend the collocation change its profile either in size or shape.

4.1 INFLUENCE OF FEATURES

The pattern design comprises of all assorted features in three types, co-features and feature set which are independent features, associative features (dependent features) and grouped features respectively. The associative features develop inevitably when the intensity of indispensability of any adjacent feature develops enormously and they come into existence with association of already existing feature. Independent features have high degree of indispensability, as they inherent and also carry core characteristics of a spatial object that retain the basic nature. Features that are alike and having same feature sets may be combined to represent a huge characteristic of a spatial object thus become grouped feature.

Every feature is said to be as member of a collocation based on its indispensability factor. However there are greater chances of changing probability for feature types defined for the formation of a collocation. The indispensability can be expressed in a degree of scale, consisting equivalent to the elements of a fuzzy set.

Probability of indispensability of a feature is

$$P(I(f)) \quad \ldots (2)$$

or

$$P(I(f_i)) \quad \ldots (2a)$$

The number of indispensable features from a pattern or a collocation pattern;

$$A = \sum_{i=1}^{n} P(I(f_i)) \quad \ldots (3)$$

Where $I(f_i) > 0$, indicates the degree of indispensability;

Probability of shift is another factor that states the indispensability. The spatial element always indicates the shift of the feature or features. More simply, the features of a spatial object change naturally, if the change insures the existence of the feature, then it must be validated with its degree of indispensability. If not existing then the feature remains unseen in that spatial object, the feature is more dispensable. If a feature f_i can change in k-ways to support its existence in the collocation, when k>0, then

$$S = P(f_{i_k} \mid 1 \le f_i \le k) \quad \ldots (4)$$

Thus equation (4) states the shift of indispensability of a feature *i*. The indispensability is an outcome of *natural experiment*. Hence, the natural experiment being a random variable, changes made to the expected feature and the expected collocation respectively is;

$$E(f(k)) \quad \ldots (5)$$

4.2 PROBABILISTIC SUPPORT

Suppose there are n collocations C_0 to C_{n-1}. Determine which collocation a feature type F is most associated with means calculating the probability that feature type F is in collocation C_i, written $P(C_i|F)$, for each collocation C, Using Bayes Rule, we can calculate $P(C_i|F)$

$$\frac{P(F/C_i)P(C_i)}{P(F)}$$

$P(C_i|F)$ is the probability the feature type F is in collocation C_i, that is the probability that given a random feature set of feature type F, they appear in collocation C_i. $P(F|C_i)$ is the probability that for a given collocation C_i, the feature type F in that collocation. $P(C_i)$ is the probability if a given collocation that is the probability of feature type being in collocation C without considering its feature set, characteristics with respect to structure of the spatial object. P(F) is the probability of that specific feature occurring. To calculate which feature type F, should be a member of collocation, $P(C_i|F)$ need to be calculated for each of the collocation and find the largest probability. Because each of those calculations involve the unknown but fixed value P(F), can be ignored and calculated as

$$P(C_i \mid F) = P(F \mid C_i) \times P(C_i)$$

When relative probability is the interest the P(F) can also be safely ignored $P(C_i|F)$ and P(F) simple acts as a scaling factor on $P(C_i|F)$. F is split into set of features, the values to intensify the features, F being a feature type. A stochastic process SF may be applied to obtain the necessary feature set for F, i.e., $f_0 \ldots f_{m-1}$. For obtaining $P(F|C_i)$, the product of the probabilities for each feature (f) of a feature type (F) is calculated.

The stochastic process is ideal to discuss in this regard. Generally a stochastic process is a random function or a random variable. In the context of spatial domain, consider each collocation referred by a lead value, and as the changes acquired to the collocation are random, suppose, for each point that refers the collocation-lead in a sample space is assigned a number, a *function* is assumed to be contained on the sample space. The *function* is a *random variable* and more precisely a *random function* defined with all physical, geometrical and other significant quantities represented by the features of collocation. If a fixed number of changes are to be analyzed in the change of collocation due to uncertain or certain changes in features then the random variable is considered to be as *discrete random variable*, otherwise a *continuous random variable*. The mathematical expectation of a discrete random variable **X** representing the possible changes of collocations $c_{1\ldots}\, c_n$, is defined as:

$$\mathbf{E(X)} = c_1\mathbf{P}(\mathbf{X}=c_1)+\ldots+c_n\mathbf{P}(\mathbf{X}=c_n)$$

Or equivalently, the probability distribution

$$\mathbf{P}(\mathbf{X}=c_i) = f(c_i)$$

The **E(X)** derives the input for the different states of the collocation, which the states may be considered into a random sequence of states. The change probability in either shape or size or dependently, is denoted by Pr(S) =*s*, where *s* is a real number such that $s \in [0,1]$. Pr(S) is derived into 1, when an uncertain event of change of size or shape will occur. Further if S is $\bigcup_{i=1}^{\infty} S_i$, for all *i* not equal to *j*, S_i and S_j are independent events, also S_i does not depend on S_j, and they are distinct, the probability is the product of all S, that means if S_i occur or not S_j occurs and there is a probability of change occurring in the spatial domain.

The directional probabilistic notation is given as:

$$P(c \mid P(f \nmid C \supseteq f.\forall \overset{n}{\underset{i=1}{P_I}}(f_i))) \quad \ldots (6)$$

The changes of collocation in the spatial series can be explained with the Markovian chain. Let a random sequence for the collocation states, ***Col (n)*** may be considered, which is called Markov-*p*, or p^{th}-order Markov, according to Markov Process of Random Signals. Let us consider the past changes made to the collocation rule as ***Col(n)***; If the conditional probability of ***Col(n)*** given the entire past is equal to the conditional probability of ***Col(n)*** given only *C(n-1),…C(n-p),* i.e.,

$$\mathbf{P}[C(n)|\, C(n-1), C(n-2),] = \mathbf{P}[C(n)|\, C(n-1)], C(n-p)] \;\forall\; n$$

According to Markov Random Fields in two dimensions, A Markov-1 sequence is simply called Markov. A Markov-*p* scalar sequence can also be expressed as a $(p \times 1)$ Markov-1 vector sequence. Another interpretation of a p^{th} –order Markov sequence is that if the "present" is known, then the "past" and the "future" are independent.

5. STATE DIAGRAM

A state diagram can be applied to represent the changing collocations. The states of collocations are irrecurrable and unpredictable they are manifested by so many natural experiments even like geographical disorders and reactions. The spatial domain is a spatial system for which the irrecurrable changes occur on it and in turn on the collocation are recorded as variable states. It is not possible to represent the continuous changing of states of collocation using conventional state diagrams. A system state diagram **Figure 2** to define to establish the schematic structure of the experiment.

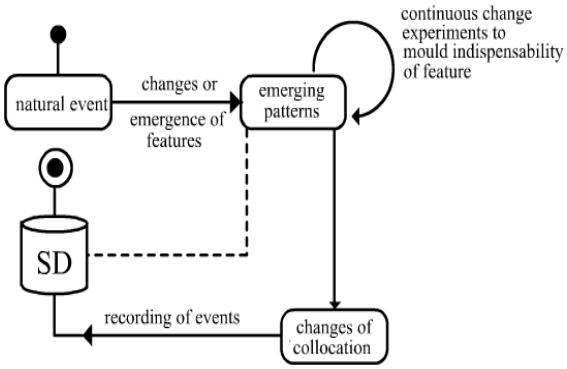

Fig. 2 System State Diagram

Types of natural events can be identified as external events, internal events and temporal events.

State: The cumulative results of the behavior of a collocation, one of the possible condition in which a collocation may exist, characterized by definite features that are distinct from other features, at any given point in time, the state of an collocation encompasses all of the (usually static) properties of features and the current (usually dynamic) values of its each of the properties.

State space: enumeration of all possible states of a collocation the state space of a collocation encompasses an indefinite yet finite number of possible (although not always desirable nor expected) states.

5.1 COLLOCATION REPRESENTATION

The collocation is a basic pattern and the collocation rule is the spatial knowledge**.** How to represent the collocations are given in **Figure 3.** Where the weighted features are coagulated as *principal* and others are connected to the principal as groups of sets called as *rank* of the principal. The feature grouped as in the rank describes the inevitability and the strong relationship of the features in the principal.

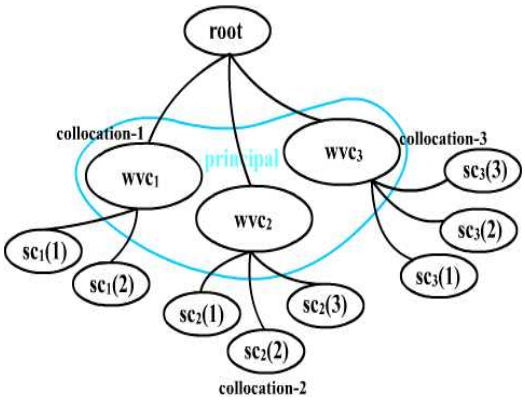

Fig.3 illustrating the A principal-ranking structure for spatial knowledge representation .

6. CONCLUSION

The probability notations are assumptions for the phenomena of change of collocation. This paper provokes study of the nature of collocation with only notational significance in probability. The representation of collocation and its allied parts are more probable in a spatial domain. The probability of occurrence of a collocation and its different varying natures are identified and denoted with several levels of probabilistic equations. Baye's rule helps in sorting out the indispensable features. The result achieved in this paper throws light on probabilistic idea that collocations change in their size and shape.

7. REFERENCES

[1] Chawla, Shekkar," Spatial Databases: A Tour", Prentice Hall, 2002.

[2]David J. Hand, "Statistics and Data Mining: Intersecting Disciplines", Department of Mathematics, Imperial College, London, UK. SIGKDD Explorations, ACM SIGKDD, June 1999.

[3] Huang, Shekkar, Xiong, "Discovery Collocation Patterns from Spatial Data Sets: A General Approach", IEEECS, 2004.

[4] M.Nagabhushana Rao , A.Ramamohan reddy, P.Govindarajulu, "Spatial Knowledge Management System Framework Through Self Adaptive Modeling ",International Journal of Computer Science and Network Security , vol.6 No.8A,August 30,2006.

[5] M.Nagabhushana Rao, A.Ramamohan reddy, P.Govindarajulu, "A Spatio - Temporal Approach to Identify Variable Size/Shape Collocation ", International Journal of Computer Science and Network Security, vol.6 No.9A, September 30,2006.

[6] Shashi Shekhar, Sanjay Chawla, Siva Ravada, Andrew Fetterer, Xuan Liu, Chang-tien Lu, "Spatial Databases: Accomplishments and Research Needs", IEEE TKDE, Vol. 11, No. 1, January, PP.45,1999.

[7] Yan Huang, Sashi Shekhar," Discovering Spatial Co-location Patterns: A Summary of Results", 7th Intl. Symposium on Spatio-Temporal Batabases, 2000.

[8] Waldo Tobler, " Global Spatial Analysis of *Computers, Environment, and Urban Systems* ", PP.493-500, Elsevier Science Ltd.,2002 .

[9] Sanjay Chawla, Florian Verhein, "Mining Spatio-Temporal Association Rules, Sources, Sinks, Stationary Regions and Thoroughfares in Object Mobility Databases", Lecture Notes, Springer Berlin/Heidelberg, Computer Science, PP.187-201, Mar-2006.

8. AUTHORS

Prof. M.Nagabhushana Rao, Professor in Department of Computer Science, C.R.Engineering Tirupati is also a Research Student in Computer Science, Sree Venkateswara University

Tirupati. Completed his BE in Computer Science from Amaravathi University, Completed MS in Software Systems from BITS, Pilani. He submitted his PhD .He is having 15 years of teaching experience. He is having three international journal publications and three International Conference Papers.

Prof. P.Govindarajulu, Professor in Department of Computer Science at S.V.University, Tirupati, has completed M.Tech in Computer Science from IIT Madras (Chennai), PhD from IIT Bombay (Mumbai).He worked in various positions Dean, B.O.S–Chairman. His area of research is Databases, Data mining, Image Processing, Software Engineering, and Software Engineering.