Volume 3, Issue 11, November 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Hierarchical Window Centric Method of Modeling Spatial Co-Location Patterns on Spatial Database

G. Kiran Kumar, K.L. Chugh, Sarita Choudhury
Dept. of CSE,
MLR Institute of Technology
Hyderabad, India

Abstract- ***With the growing technology, most of the focus is on spatial data mining. Spatial data mining refers to the extraction of unknown and unexpected information from spatial data sets of massive, high dimensionality and complex spatial databases. Spatial data mining is a process to discover related knowledge, potentially constructive and high utility patterns embedded in geographic Information. Application specific tools for extracting efficient and useful information from spatial data sets can be of great importance to the organizations which own, generate and manage large databases. The objective of co-location pattern mining is to find the subset of features frequently located together in the same region. Three modeling methods for co-location patterns[16,19] are event centric model, feature centric model and window centric model. We have proposed a new method for modeling co-location patterns "Hierarchical Window Centric Model", which is an extension of Window Centric Model. We would carry out experimental evaluations and performance tuning in the near future.***

Keywords: Spatial data Mining, Co-location rules, window centric model

1. Introduction

The fast growing progress in computerized data acquisition and storage results in the growth of enormous databases. With the continuous increase and accumulation, the huge amounts of the computerized data have far exceeded human ability to completely interpret and use. These phenomena may be more serious in geo-spatial science. In order to understand and make full use of these data repositories, a few techniques have been tried, e.g. expert system, database management system, spatial data analysis, machine learning, artificial intelligence and environmental studies[25,27]. In 1989, knowledge discovery in databases was further proposed. In 1995, data mining also appeared. As both data mining and knowledge discovery in databases virtually point to the same techniques, people would like to call them together, i.e. data mining and knowledge discovery (DMKD)[1]. As 80% data are geo-referenced, the necessity forces people to consider spatial characteristics in DMKD and to further develop a branch in geo-spatial science. The Earlier research of data mining concentrated on non-spatial data, after requirements in business like insurance, finance, etc. were fulfilled, emphasis on research has been consequently shifted from non-spatial data (market basket data) to spatial data[20,22]. There is an observable and important difference between non-spatial data and spatial data: all spatial datum are governed by scales in nature. A spatial object can be characterized accurately and correctly, by stating the exact concomitant scale of the spatial objects.

In general, spatial data mining, or knowledge discovery is the process of discovering implicit knowledge and previously unknown, but potentially useful patterns from large spatial datasets that are not explicitly represented in the databases[17,18]. These techniques can play an important role in understanding spatial data and in capturing intrinsic relationships between spatial and non-spatial data. Moreover, such discovered relationships can be used to present data in a concise manner and to reorganize spatial databases to accommodate data semantics and achieve high performance. Spatial data mining has wide applications in many fields, including GIS systems, image database exploration, medical imaging, etc.[2-7]. The amount of spatial data obtained from satellite, medical imagery and other sources has been growing tremendously in recent years. A crucial challenge in spatial data mining is the efficiency of spatial data mining algorithms due to the often huge amount of spatial data and the complexity of spatial data types and spatial accessing methods. Nowadays, large amount of spatial data have been collected from many applications and data collection tools. "The spatial data explodes but knowledge is poor" [8], therefore, "We are drowning in data, but starving for knowledge!" The implicit knowledge hidden in those spatial data cannot be extracted using traditional database management systems.

The growing production of maps is generating huge volumes of data that exceed people's capacity to analyze them. It thus seems appropriate to apply knowledge discovery methods like data mining to spatial data. This recent technology is an extension of the data mining applied to alphanumerical data on spatial data[26,28]. The main difference is that spatial analysis must take into account spatial relations between objects. The applications covered by spatial data mining are decisional ones, such as geo-marketing, environmental studies, risk analysis and so on. For example, in geo-marketing, a store can establish its trade area, i.e. the spatial extent of its customers and then analyze the profile of those customers on the basis of both their properties and the properties related to the area where they live. Nowadays, data

analysis in geography is essentially based on traditional statistics and multidimensional data analysis and does not take into account of spatial data [9]. Yet the main specificity of geographic data is that observations located near to one another in space tend to share similar (or correlated) attribute values. This constitutes the fundamental of a distinct scientific area called "spatial statistics" which, unlike traditional statistics, supposes inter-dependence of nearby observations. An abundant bibliography exists in this area, including well-known geo-statistics, recent developments in Exploratory Spatial Data Analysis (ESDA) by Anselin and Geographical Analysis Machine (GAM) by Openshaw. Multi-dimensional analytical methods have been extended to support contiguity [10]. We maintain that spatial statistics is a part of spatial data mining, since it provides data-driven analysis. Some of those methods are now implemented in operational GIS or analysis tools. This paper is organized as follows. We first discuss about Spatial Data Mining in section 2. Section 3 discusses about related work. In section 4 actual work that is modeling of Co-location patterns are discussed and finally we offer our conclusion & future work in Section 5.

2. Spatial Data Mining

The spatial data mining is a newly developed edge course. When computer, database, management decision support technique and etc. are applied, then certain stage Spatial data are more complex, more changeable and bigger that common affair datasets [1]. Spatial dimension means each item of data has a spatial reference where each entity occurs on the continuous surface, or where the spatial-referenced relationship exists between two neighbor entities. Spatial data includes not only positional data and attribute data, but also spatial relationships among spatial entities. Moreover, spatial data structure is more complex than the tables in ordinary relational database. Besides tabular data , maps have been widely used as the main references in the playing field of geography. There are two ways to represent thematic maps: raster and vector. In the raster image form thematic maps have pixels associated with the attribute values, these are the graphic data in spatial database and the features of graphic data are not explicitly stored in the database. At the same time, using GIS, the user can query spatial data and perform simple analytical tasks using programs or queries. GIS have only basic analysis functionalities, the results of which are explicit. It is under the assumption of dependency and on the basis of the sampled data that geo-statistics estimates at un-sampled locations or make a map of the attribute. Because the discovered spatial knowledge can support and improve spatial data-referenced decision-making, a growing attention has been paid to the study, development and application of SDMKD [11].

Spatial data mining has become an important research area in order to analyze very large spatial databases. Among the better known spatial data mining approaches, we find the generalization based methods, clustering, spatial associations, approximation and aggregation, mining in image and raster databases, spatial classification, and spatial trend detection. However, these approaches do not consider all the elements found in a spatial database (spatial data, non-spatial data and spatial relations among the spatial objects) in an extended way. Some of them focus first on spatial data and then on the non-spatial data or vice versa, and others consider restricted combinations of these elements. We argue that if we are able to mine them as a whole and not as separated elements (because they are related elements) we could find patterns that might contain both types of data and spatial relationships enhancing the quality of the results (being more descriptive). A graph based representation provides the flexibility to describe.

Spatial data mining is the process of finding out interesting and previously unknown, but possibly useful patterns from large spatial database[12,13]. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation.

3.0 Related Work

3.1 Modeling of Co-Location Patterns. This section defines approaches to model co-location rules as a substitute to create explicit disjoint transactions from continuous spatial data.

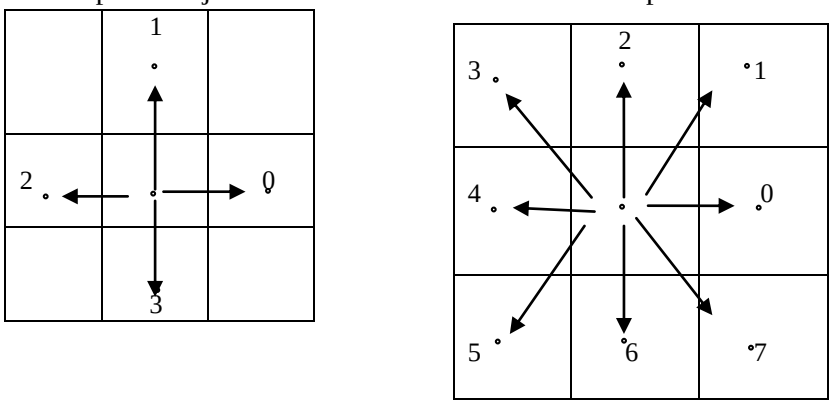

Fig 3.1 spatial data set illustrating Neighborhood having 4 and 8 connectivity

In Figure 3.1 a uniform grid is imposed on the underlying spatial framework. For each grid *l*, it has two horizontal and two vertical neighbors making 4 neighbors and the connectivity is called 4-connectivity, similarly by adding four diagonal neighbors we get totally eight neighbors and the connectivity is called 8- connectivity.

We have three types of modeling co-location rules. They are Reference Feature Centric Model, Window Centric Model and Event Centric Model[14,15,22].

3.2 Reference Feature Centric Model

This model is applicable to application domains focusing on a specific Boolean spatial feature, e.g. cancer. This model computes neighborhoods to "materialize" a set of transactions around instances of the reference spatial feature.

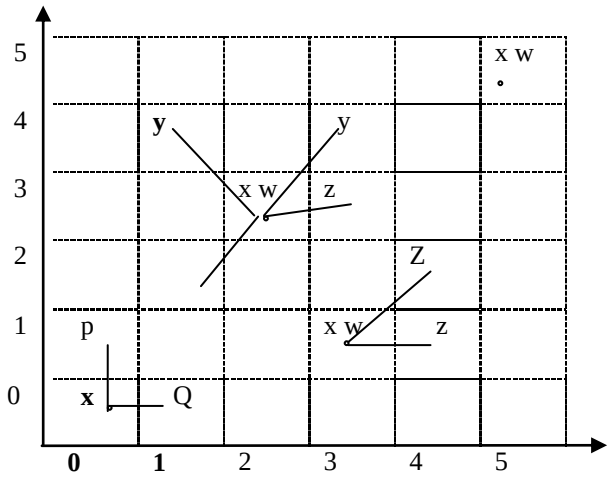

Fig. 3.2 Reference Feature Centric model

With the help of "materialized" transactions the support and confidence of the traditional association rule problem may be used as a prevalence and conditional probability.

In reference feature centric model, the instances of 'X 'are connected with their neighboring instances of 'Y' and 'Z' as shown in figure 3.2. The set of spatial predicates include one predicate named "near-by". We define near-by(x,y) to be true if and only if y is x's neighbour.

Then for each instance of spatial feature X, a transaction which is a subset of relevant features {Y, Z} is defined. For example, for the instance of X at (2, 3), transaction {Y, Z} is defined because the instance of Y at (1, 4) (and at (3, 4)) and instance of Z at (1, 2) (and at (3, 3)) are close to (2, 3). The transactions defined around instances of feature 'X' are summarized in table 3.1

Table 3.1 Transactions of feature 'X' given in Figure 3.2

Instance of X	Transaction
(0,0)	{P,Q}
(2,3)	{Y,Z,P}
(3,1)	{Z}
(5,5)	¢

3.3 Window Centric Model

This model is applied to applications like mining, surveying and geology, which focus on land-parcels. A goal is to calculate sets of spatial features likely to be discovered in a land parcel given that some other features have been found there. The window centric model enumerates all possible windows as transactions. In a space discretized by a uniform grid, windows of size k X k can be enumerated and materialized, ignoring the boundary effect [23]. Each transaction contains a subset of spatial features of which at least one instance occurs in the corresponding window. The support and confidence of the traditional association rule problem may again be used a prevalence and conditional probability measures. There are sixteen 3X3 windows corresponding to 16 transactions in Fig 3.3. All of them contain X and 15 of them contain both X and Y. An example of an association rule of this model is: an instance of type X in a window → an instance of type Y in this window with 15/16 = 93.75% probability. A special case of the window centric model relates to the case when windows are spatially disjoint and form a partition of space. This case is relevant when analyzing spatial datasets related to the units of political or administrative boundaries (e.g. country, state, zip-code). In some sense this is a local model since we treat each arbitrary partition as a transaction to derive co-location patterns without considering any patterns across partition boundaries. The window centric model "materializes" transactions in a different way from the reference feature centric model[8].

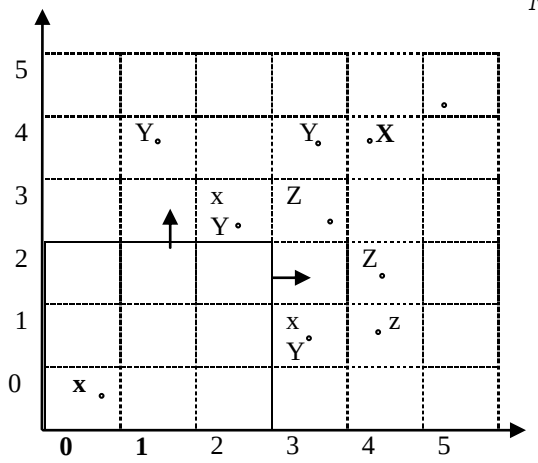

Fig. 3.3. Window Centric model

3.4 Event Centric Model

This model is applied to applications like ecology where there are many types of Boolean spatial features. Ecologists are interested in finding subsets of spatial features likely to occur in a neighborhood around instances of given subsets of event types. For example, let us determine the probability of finding at least one instance of feature type Y in the neighborhood of an instance of feature type X in Fig 3.4. There are three instances of type X and only one of them have some instance(s) of type Y in their 9-neighbor adjacent neighborhoods [24]. The conditional probability for the co-location rule is: spatial feature X at location l → spatial feature type Y in 9-neighborhood is 25%. Neighborhood is an important concept in the event centric model.

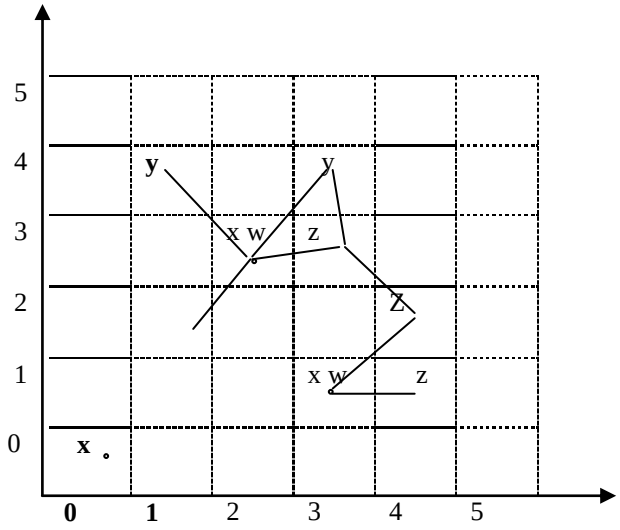

Fig. 3.4. Event Centric model

4.0 Proposed model: Hierarchical window centric model

It is an extension of window centric model for modeling co-location patterns where spatial data is first focused on certain set of spatial features and it is divided into four parts(windows), and at each part window centric model can be applied. We carried out experimental results with the above concept and we found the better co-location patterns than window centric model . Hierarchal window centric model is mainly aimed at identifying the frequent patterns from the mushroom dataset using different association rule mining algorithms. Details are as follows:

In this example, window centric model is converted into transaction table, we are taking 5 item sets and 5 transactions at each time and apply apriori algorithm each time

There are nine 3x3 windows in each windows (W_1,W_2,W_3 ,W_4) clockwise corresponding to 36 transactions in Fig 4.1. In W_1, all 9 of them contain X and all 9 of them contain both X and Y. An example of an association rule of this model is: in W1 an instance of type X in a window → an instance of type Y in this window with 9/9 = 100% probability. In W_2, all of them contain X and 7 of them contain both X and Y. An example of an association rule of this model is: in W_2 instances of type X in a window → an instance of type Y in this window with 7/9 = 77.7% probability.

In W_3, 7 of them contain X and 4 of them contain both X and Y. 2 of them contain only Y. An example of an association rule of this model is: in W_3 instances of type X in a window → an instance of type Y in this window with 6/9 = 66.6% probability.

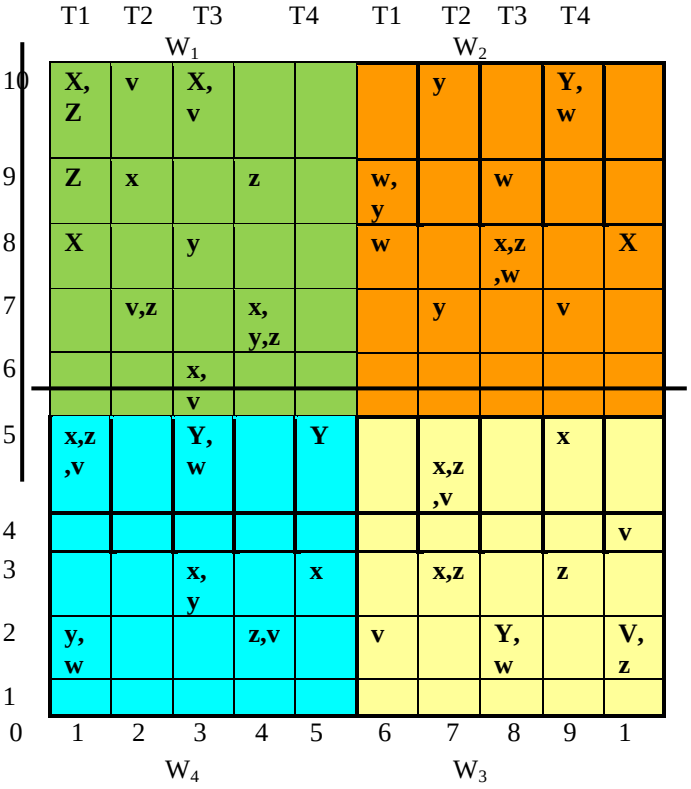

Fig.4.1

In W_4, all of them contain X and all of them contain both X and Y. An example of an association rule of this model is: in W_4 instances of type X in a window → an instance of type Y in this window with 9/9 = 100% probability.

5. Conclusion and Future Work

The proposed model "Hierarchical Window Centric Model" gives better co-location patterns than other existing models such as Reference Feature centric model, window centric model and event centric model. In future, the same concept can be extended by applying event centric model in each window.

References:

1. Deren LI and Shuliang WANG, " Concepts, principles and applications of spatial data mining and knowledge discovery", in ISSTM 2005, August, 27-29, 2005, Beijing, China .
2. W. Wang, J. Yang, and R. Muntz. "STING: A statistical information grid approach to spatial data mining" in International Conference on Very Large Data Bases, Athens, Greece, Morgan Kaufman, San Mateo, CA, pp. 186–195, 1997.
3. M. S. Chen, J. Han, P. S. Yu. Data mining, an overview from database perspective *IEEE Transactions on Knowledge and data Engineering*, 1997.
4. U. Fayyad, G. P.-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, Vol. 17 No. 3, pp. 37-54, Fall 1996.
5. U. Fayyad, G.P.Shapiro, P.Smyth, and R.uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining*. AAAI /MIT Press, Menlo Park, CA, 1996.
6. K. Koperski, J. Adhikary, and J. Han. Spatial data mining: progress and challenges. *SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery* (*DMKD'96*), Montreal, Canada, June 1996.
7. K. Koperski and J. Han. Data mining methods for the analysis of large geographic databases. *Proc. 10th Annual Conf. on GIS*. Vancouver, Canada, March 1996
8. G Kiran Kumar, P Premchand, T V gopal, "A Novel Method of Modeling Spatial Co-Location patterns on Spatial Database", ICoFCS-2011, IISc, Banglore.
9. Li D.R., Wang S.L., Li D.Y. and Wang X.Z., 2002, Theories and technologies of spatial data knowledge discovery. Geomatics and Information Science of Wuhan University 27(3), 221-233.
10. K Zeitouni "A survey of spatial data mining methods databases and statistics point of views", Data warehousing and web engineering, 2002 - books.google.com.

11. G Kiran Kumar, P Premchand, "Extraction of High Prevalence based Co-location Pattern on Spatial Database Using a Combinatorial Approach", Journal of Data engineering and computer science-2012
12. Longley P. A., Goodchild M. F., Maguire D. J., Rhind D. W., Geographical Information Systems - Principles and Technical Issues, John Wiley & Sons, Inc., Second Edition, 1999.
13. Han, J., Kamber, M., 2001, Data Mining: Concepts and Techniques (San Francisco: Academic Press)
14. S Shekhar, P Zhang. Data Mining and Knowledge Discovery 2010 – Springer .
15. G Kiran Kumar, P Premchand, T V gopal, "Mining Of Spatial Co-location Pattern from Spatial Datasets" , International Journal of computer applications-2012, Volume 42,Serial No:21.
16. R. Agarwal and R. Srikant. "Fast algorithms for mining association rules," in Proc. of the 20th Int'l Conference on Very Large Data Bases, Santiago, Chile, pp. 487–499, 1994.
17. X. Zhang, N. Mamoulis, D.W.L Cheung, and Y. Shou. "Fast mining of spatial collocations," in *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, pp. 384–393, 2004.
18. Y. Morimoto. "Mining frequent neighboring class sets in spatial databases," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 353– 358, 2001.
19. V. Estivill-Castro, and I. Lee. "Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data", in Proc. of the 6th International Conference on Geo-computation, pp. 24–26, 2001.
20. V. Estivill-Castro, and A. Murray. "Discovering associations in spatial Data—an efficient medoid based approach," in Proc. of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Berlin Heidelberg New York, pp. 110–121, 1998.
21. R. Munro, S.Chawla, P. Sun. "Complex spatial relationships," The Third IEEE International Conference on Data Mining (ICDM2003), IEEE Computer Society, p. 227, 2003.
22. N.A.C. Cressie. Statistics for Spatial Data. Wiley: New York ISBN:0471843369, 1991.
23. Y. Chou. Exploring Spatial Analysis in Geographic Information System. Onward Press: Santa Fe, NM ISBN:1566901197, 1997.
24. S. Shekhar, and Y. Huang. "Co-location rules mining: A summary of results," in *Proc. 7th Intl.Symposium on Spatio-temporal Databases*, Springer, Berlin Heidelberg New York, p.236, 2001.
25. K. Koperski, and J. Han. "Discovery of spatial association rules in geographic information databases," in *Proc. of the 4th International Symposium on Spatial Databases*, Springer, Berlin Heidelberg New York, pp. 47–66, 1995.
26. Yan Huang, Jian Pei, Hui Xiong, " Mining Co-Location Patterns with Rare Events from Spatial Data Sets Geoinformatica (2006) 10: 239–260.
27. G Kiran Kumar, P Premchand, "A Novel Hybrid Spatial Clustering Algorithm", International Journal of Engineering Research and Applications, Volume 2 Issue 3,May-Jun 2012
28. Dr.M.Hemalatha, N. Naga Saranya,"A Recent Survey on Knowledge Discovery in Spatial Data Mining" International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011 ,ISSN (Online): 1694-0814.

