Spatial Co-location Patterns Mining

Ruhi Nehri
Dept. of Computer Science and Engineering.
Government College of Engineering, Aurangabad,
Maharashtra, India.

Meghana Nagori
Dept. of Computer Science and Engineering.
Government College of Engineering, Aurangabad,
Maharashtra, India

ABSTRACT

Data mining refers to a process of analyzing data from different perspectives and summarizing it into useful information that can be used in variety of data centric applications in real time. Geographical Information System (GIS) combined with Data Mining has long being an area of research. GIS is a system that provides information about spatial data. Spatial Data consist of data about real world such as a point on a map represent a place given its latitude and longitude information. Such information of real world constitutes spatial datasets. Data mining in GIS provides a way to analyze this spatial datasets to provide a desired result. Knowledge discovery is a process of extracting implicit knowledge or information from spatial data. Co-location patterns discovery refers to a process of finding subsets of spatial features that are located in close proximity. Spatial Co-location pattern discovery is finding coexistence of non-spatial features in a spatial neighborhood. In this paper we have mined co-location patterns with an approach based on participation index and participation ratio. This technique finds the maximal participation index and uses a clustering algorithm. We have used aprior algorithm method that yields better performance improvement to the algorithm used.

General Terms

Spatial data, Co-location pattern mining.

Keywords

GIS, Data Mining, Spatial Data Mining, Co-location Patterns, Participation index, participation ratio

1. INTRODUCTION

1.1 Geographical Information System

Geographical Information System (GIS) refers to a system that capture, store, manipulate, analyze, manage and present geographical data. This geographical data is usually referred as spatial data. A GIS is a system that digitally makes and manipulates spatial areas. GIS can be also termed as spatial data mining. GIS or spatial data mining is an application of data mining methods to spatial data. Data mining offers a great potential benefit when applied to GIS based decision making applications such as environmental studies. A characteristic of such applications is that spatial correlation between data measurements requires the use of special algorithms for more efficient data analysis.

1.2 Data Mining

Data mining can be defined as a process of analyzing data from different perspective and summarizing it into useful information. This mined information can be used in many applications such as increase revenue, cut, cost or both. Data mining is also called as data or knowledge discovery. More technically, data mining is the process of finding correlations or useful patterns among large relational databases. Data mining, properly known as Knowledge Discovery in Databases (KDD), is the nontrivial extraction of implicit, previously unknown and potentially useful information for data in databases [1]. Data mining is actually a process of finding hidden information/patterns from large repositories of data.

1.3 Spatial Data and Spatial Databases

Spatial data mining works on spatial data. Spatial data is geographical data where the underlying frame of reference is the Earth's surface [2]. Spatial data can be obtained from number of sources such as satellite images, medical equipments, video cameras, etc.

Geographical Information System (GIS) is the principle technology that motivated the interest in Spatial Database Management System (SDBMS). The main constraint while dealing with spatial data manipulating is that it is embedded in spatial database that cannot be retrieved using conventional database management systems. It needs an advance database management system that tackles all its operations. This need was recognized by scientist, administrators and environmental researchers. Thus a Spatial Database Management System is used for this purpose. GIS provides mechanism for analyzing, manipulating and presenting geographic data. GIS can apply set of operations over few objects and layers, whereas SDMS can apply operations over a set of objects and set of layers. For example, a GIS can list of the neighboring countries of a given country (e.g., India). However, it is a tedious work in GIS to perform set operations or queries like, list the countries with highest number of neighboring countries. Set- based operations or queries can be well performed in SDBMS. SDBMS aims at effective and efficient management of spatial data related to space, engineering designs, conceptual information space (multidimensional decision support system), and etc [2] .In this paper, section 2 describes spatial data mining. Section 3 describes co-location pattern mining. Section 3 also contains a case example of co-location pattern mining and its algorithm. The proposed solution to mine co-location patterns is described in section 4. Conclusion and future scope is given in section 5.

2. SPATIAL DATA MINING

Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases [8]. Usually spatial data is in huge amount (in terabytes). Thus making it costly and unrealistic for users to examine such spatial data in detail. Spatial data mining is a field aimed at automating such a knowledge discovery process [8]. Thus, it plays a very important rule to capture intrinsic relationship between spatial and non-spatial data, extracting interesting spatial patterns, presenting data at a higher conceptual level.

Spatial data mining is a special kind of data mining. The main difference between data mining and spatial data mining is that in spatial data mining tasks we not only use non-spatial attributes but also spatial attributes. Non-spatial attributes refers to a specific, precisely defined location. It is independent of all geometric considerations, for example, a person's height, age, etc. Spatial attributes refers to data that defines a specific location. It includes location, shape, size and orientation, for example, latitude and longitude information of a specific location. Spatial data includes spatial relationships. Different spatial data mining algorithms have been developed for knowledge discovery from spatial data. Some of the basic spatial data mining tasks are spatial classification, association rules, clustering and trend detection.

2.1 Spatial Classification

Classification of objects in data mining is based on the concept of assigning an object to a class from a given set of classes according to its attribute. Finding a set of rules determining the class of an object based on it attributes values. Classification is a type of "Supervised learning". Spatial classification is the process of predicting class labels based on the characteristics of entities as well as the spatial relationships to other entities and its characteristics [7].

2.2 Spatial Association Rules

Association rules were discovered to find regularities between items in a large relational database. Similarly as association rules are mined in relational or transactional databases, spatial association rules can be mined in spatial databases considering the entities spatial properties and predicates. Spatial co-location pattern mining is conceptual very similar, but technically different from association rule mining. A co-location pattern represents subsets of features that are frequently located together. Of course, a location is not a transaction and two features rarely exist at exactly the same location. Thus, a user-specified neighborhood is needed as a container to check which features co-locate in the same neighborhood [3]. Many Algorithms has been proposed to mine spatial co-location patterns.

2.3 Spatial Clustering

Clustering refers to a process of creating group of data organized at some similarity among the members of the dataset. The aim of clustering is to group the objects from database into clusters in such a way that objects of similar characteristics forms one cluster and objects from different cluster are dissimilar Clustering methods can be broadly classified into two groups: partitioning clustering and hierarchical clustering. Three types of clustering methods has been studied to consider spatial information in clustering which includes spatial clustering, regionalization and point pattern analysis [3].

2.4 Trend Detection

A trend can be defined as a temporal pattern recognized in some time series data. Trend detection in spatial data mining finds trends in database. Spatial trend is defined as a pattern of change of a non-spatial attribute in the neighborhood of a spatial object.

3. CO-LOCATION PATTERNS

Spatial co-location patterns can be mined as similar to association rules. Spatial association rules can be mined in spatial databases by considering spatial properties and predicates. A spatial association rule is expressed in the form A => B[s%, c%] where A and B are sets of spatial or non-spatial predicates, s% is the support of the rule and c% is the confidence of the rule [3].

3.1 Spatial Co-location Pattern Mining

Spatial association rule uses spatial predicates like close_to, intersect, overlay, etc. it is computationally very expensive to derive spatial association rules from large spatial datasets. Thus, it is necessary to use another technique to derive rules from spatial datasets. Spatial co-location pattern mining is very similar, but technically different from association rule mining.

A Co-location is a subset of Boolean spatial features.

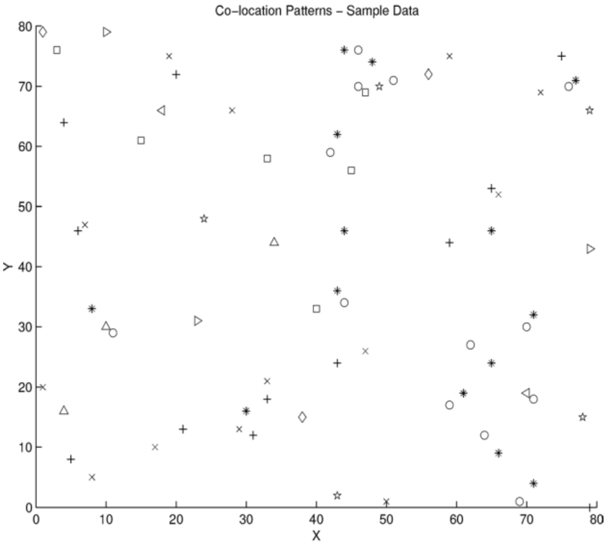


Figure 1: Co-location Pattern-Sample Data

Co-location patterns are subsets of spatial features that are located in close geographic proximity [4]. For example, the analysis of an ecology dataset may reveal symbolic species. Figure 1 shows a data set consisting of instances of several Boolean spatial features, each representing a distinct shape. A careful review reveals two co-location patterns, i.e., {'+', 'x'} and {'o', '*'}. Real-world examples of co-location patterns include symbiotic species, e.g., the Nile Crocodile and Egyptian Plover in ecology. These spatial features describe the presence or absence of geographic object types which are located at different locations. For example, road types, animal species, etc. represents Boolean spatial features.

3.2 Co-location Rules Modeling

Co-location rules are models to find the presence of spatial features in the neighborhood of instances of other spatial features. A Co-location rule is of the form: L1 → L2 (p, cp) where L1 and L2 are co-locations, p is prevalence measure and cp is the conditional probability [5]. For example, "Nile Crocodiles → Egyptian Plover" predicts the presence of Egyptian Plover birds in areas with Nile Crocodiles. The neighboring relation R is an input and can be defined using graph theory such as connected, adjacent relationship using different distance measures. The distance measures that can be used are Euclidean distance, City block distance, etc. Discovery of Co-location patterns is based on participation index and maximal participation ratio.

3.3 Co-location Mining Algorithm

In this section, co-location pattern mining algorithm is introduced. In this algorithm the prevalence measure used is the participation index. According to the algorithm, a co-location pattern is prevalent if the value of its participation index is above a user specified threshold.

Algorithm:

Step 1: Input a spatial dataset

Step2: Initialize prevalent co-location set of size 1 along with table instance

Step 3: Generate co location rules of size 2

Step 4: For co-location of size (2, 3, 4 …K–1) do

Begin

Step 5: Generate candidate prevalent co-location

Step 6: Generate table instance

Step 7: Prune co-locations using prevalence threshold

Step 8: Generate co-locations rules

Step 9: Stop

Explanation Of Step 5: Candidate prevalent co-locations of size K+1 are generated using combinatorial approach from prevalent co-locations of size K along with their table instances.

Example Case

Consider a spatial data set with spatial feature set F = {A,B, C,D,E} as shown in the figure 2. Each instance of the spatial feature is uniquely identified by its instance-id. There are 20 instances in the database.

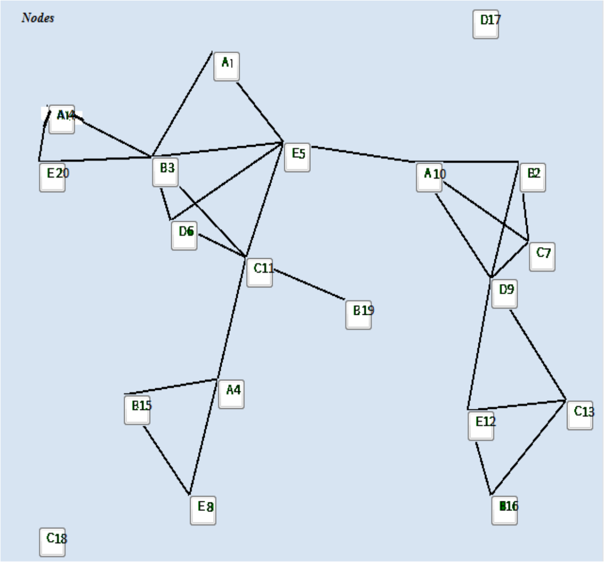


Figure 2: An Example data set

Object A has 4 instances with instance-ids { 1,4,10,14}, Object B has 5 instances with instance-ids {2,3,15,16,19}, Object C has 4 instances with instance-ids {7,11,13,18}, Object D has 3 instances with instance-ids {6,9,17}, and Object E has 4 instances with instance-ids {5,8,12,20}. The neighborhood relationship R between objects is defined based upon Euclidean distance. R can be described as two instances are neighbors if their Euclidean distance is less than a user specified threshold. These neighboring instances are connected by edges. A Clique in an undirected graph is a subset of its vertices such that every two vertices in the subset are connected by an edge. In fig.2, instance {1,3,6,5 } , {16,13,9,12} and {10,2,7,9,12} are neighbor-sets as each set forms a clique.

{A,B,C,D,E} is a co-location pattern. The neighborhood-set {10,2,7,9,12} is a row instance but the neighborhood-set {10,2,7,9,12,13} is not a row-instance of the co-location {A,B,C,D,E} because it has a proper subset {10,2,7,9,12} which contains all the features of {A,B,C,D,E}.

3.3.1 Conditional Probability

The objective of co-location pattern mining is to find frequently co-located subsets of spatial features [6]. For a co-location rule R: A → B, the conditional probability *cp(*R) of R is expressed in the form

$$\frac{|\{L \in rowset(A) | \exists \text{ L's.t. } (L \subseteq L') \wedge (L\_ \in rowset(A \cup B))\}|}{|rowset(A)|}$$

Conditional probability is the probability that a neighbor-set in *rowset*(A) is part of a neighbor-set in rowset (A ∪ B). *rowset*({A,B,C,D,E}) contains eight patterns- {12,9,2,10,13}, {12,9,16,10,13}, {5,6,3,10,11}, {5,6,3,14,11}, {5,6,3,1,11}, {12,9,2,10,7}, {12,9,16,10,7}, and {5,6,3,10,7}. Whereas r*owset*({A,B,C}) contains four patterns- {15,4,11}, {3,14,11},{3,1,11} and {2,10,7}. *rowset*({A,B,C}) = 4 out of which only 3 rows of {A,B,C} satisfy the subset condition of {A,B,C,D,E}. Thus the conditional probability *cp* ({A,B,C}, {D,E}) = 3/4=75%.

Consider an example of the co-location pattern {*traffic jam, police, road accident*} means that a traffic jam, police and a road accident frequently occur in the nearby region. From the above rowset, let us consider the occurrence of objects A,B and C represent the traffic jam and the occurrence of objects D and E as violation of road rules. When all the objects occur simultaneously, it represents the occurrence of road accident. Thus *cp* ({A, B, C} → {D, E}) represents the conditional probability of occurrence of road accident whenever a traffic is jammed i.e. it is 75% chances of a road accident.

3.3.2 Participation Ratio

The participation ratio of a spatial feature can be defined as a measure to find how a spatial feature f is co-located with other spatial features in the co-location pattern L, given a spatial database S [5]. The participation ratio Pr for a feature type fi of a co-location L if the fraction of instances of fi which participate in any row instance of co-location L. The participation ratio *pr* can be expressed in the form - *pr (* L*, fi) = π (table instance(L)) / (instances(fi)),* where L = {*f1,f2,……..,fk}.*

There exist eight patterns for the *rowset*({A,B,C,D,E}). Among all the four instances of A – (1,4,10,14) three of the instances i.e. , {1,10,14} has instances of objects B,C,D and E in their neighborhood. So the participation ratio of A is *pr*({A,B,C,D,E}, A) =3/4 Similarly *pr*({A,B,C,D,E}, B) = 3/5, *pr*({A,B,C,D,E}, C) =3/4, *pr*({A,B,C,D,E}, D) = 2/3 and *pr*({A,B,C,D,E}, E) = 1/2.

3.3.3 Participation Index

The participation index can be defined as the minimum participation ratio of instances of co-location patterns. It can be expressed in the form- $pi\ (L) = min^{k}_{i=1}$ Pr(L, *fi)* [5]. Participation index (PI) of the event is the probability of occurring of an event based on the participation ratio of the objects in the event. An event will occur only when all the objects in the event occur simultaneously. In the example, a road accident will occur only when both the traffic jam and violation of traffic rules occur simultaneously. Thus PI is the minimal of participation ratio of all the objects in the event. Hence, PI (A,B,C,D,E) of co-location pattern {A,B,C,D,E} is 1/2 which correspond to spatial feature E.

3.3.4 Maximum Participation Index

Max participation index (*Max(PI)*) represents the maximal participation ratio of the objects occurred in the event. This helps us to know that which object has influenced more in the occurrence of the event, which will help to take action. The *pr* of A,B,C,D and E are 3/4, 3/5, 3/4, 2/3 and 1/2 respectively in the occurred event. Among these, objects A and C have a higher influence of 75% in the occurrence of the event.

Consider each feature as- A as Car, D & E involve in violation of traffic rules in which D represents traffic police and E represents traffic signal. Thus, among four cars available, there is a higher probability that three cars are involved in road accident. This means 75% car are involved in road accidents. Hence, necessary action needs to be taken on the cars to reduce the risk of road accidents.

3.3.5 Unique Count

Unique Count represents count of instances of an object in a particular rule. It is derived on the basis of unique instance row-set generated of a collocation rule. For example, collocation rule for objects DBA is ({6,3,1},{6,3,14},{9,2,10}). Unique count generated for this rule is {2,2,3} i.e. the objects instances that have participated in the collocation rule are 2 instances of object D, 2 instances of object B and 3 instances of object A.

3.3.6 Medoid Participation Index

Max PI is used to figure out the most promising feature or the one that is to be prevented. In our example, the resultant features are A & C. the results shows the occurrence of cars as the prevalence measure to reduce accidents. But it is not possible to district the cars to travel on road. In this the MaxPI concept fails.

So, we propose a new concept called as MedoidPI. In this medoid of participation ratio is calculated to take decisions. The steps to calculate the MedoidPI is: 1) Calculate the Mean (M) of participation ratio. 2) Calculate the distance of each ratio from M with the help of the formula | M – x/y |, where x/y is any ratio of PR. 3) Find the minimum of the calculated ratio that will give MedoidPI. In our above example, the participation ratio of A, B, C, D and E are 3/4, 3/5, 3/4, 2/3 and 1/2 respectively. MedoidPI is 2/3 which is of feature D i.e. traffic police.

4. PROPOSED SOLUTION

In this paper, participation ratio (PR), participation index (PI) and Maximum Participation Index (MaxPI) and Medoid PI have been evaluated. The solution is based on a synthetic data set as shown in fig.1. The Co-location Pattern Mining algorithm has been designed using Visual Studio .NET 2010 environment. The programming language used to program the algorithm is C#.NET and the database used to store the spatial data information is Microsoft SQL Server 2008.

The spatial data is stored in a database table as the values <InstanceID, Instance, xcord, ycord >. InstanceID represents IDs of instances A, B, C, D, and E in the dataset. xcord and ycord represents the X-cordinate and Y-cordinate of the features A, B, C, D, and E.

Table 1: Spatial Dataset

InstanceID	Instance	Xcord	ycord
1	A	225	250
17	D	500	210
14	A	50	300
3	B	160	350
5	E	300	336
20	E	40	355
10	A	440	355
2	B	550	355
6	D	180	410
11	C	260	435
19	B	365	485
9	D	520	465
7	C	560	430
12	E	495	590
15	B	130	575
4	A	230	560
13	C	600	590
16	B	520	670
8	E	200	570
18	C	40	700

The input of spatial data is taken from Table 1. After the input, network paths are evaluated. Then the co-location rules are formulated that displays various co-locations, row-set, participation ratio, participation index (PI), Maximum PI, Medoid PI as shown in figure 3.the last rowset i.e. EDCBA in figure 3, gives co-location pattern. After generating the final row-set, the co-location patterns is drawn as shown in figure 4.

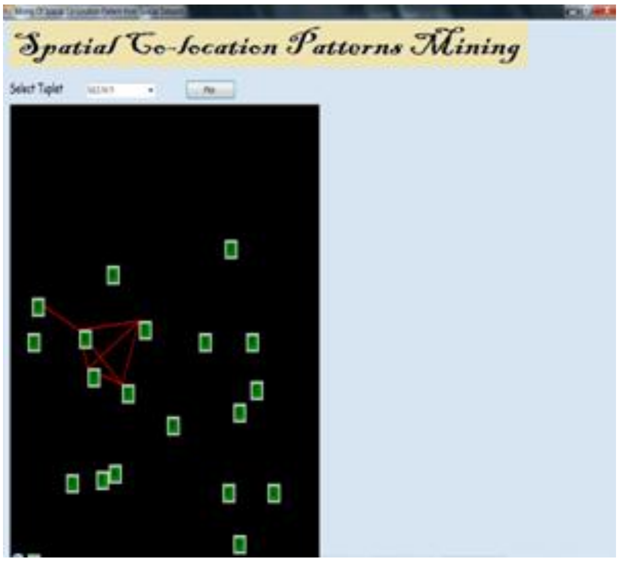
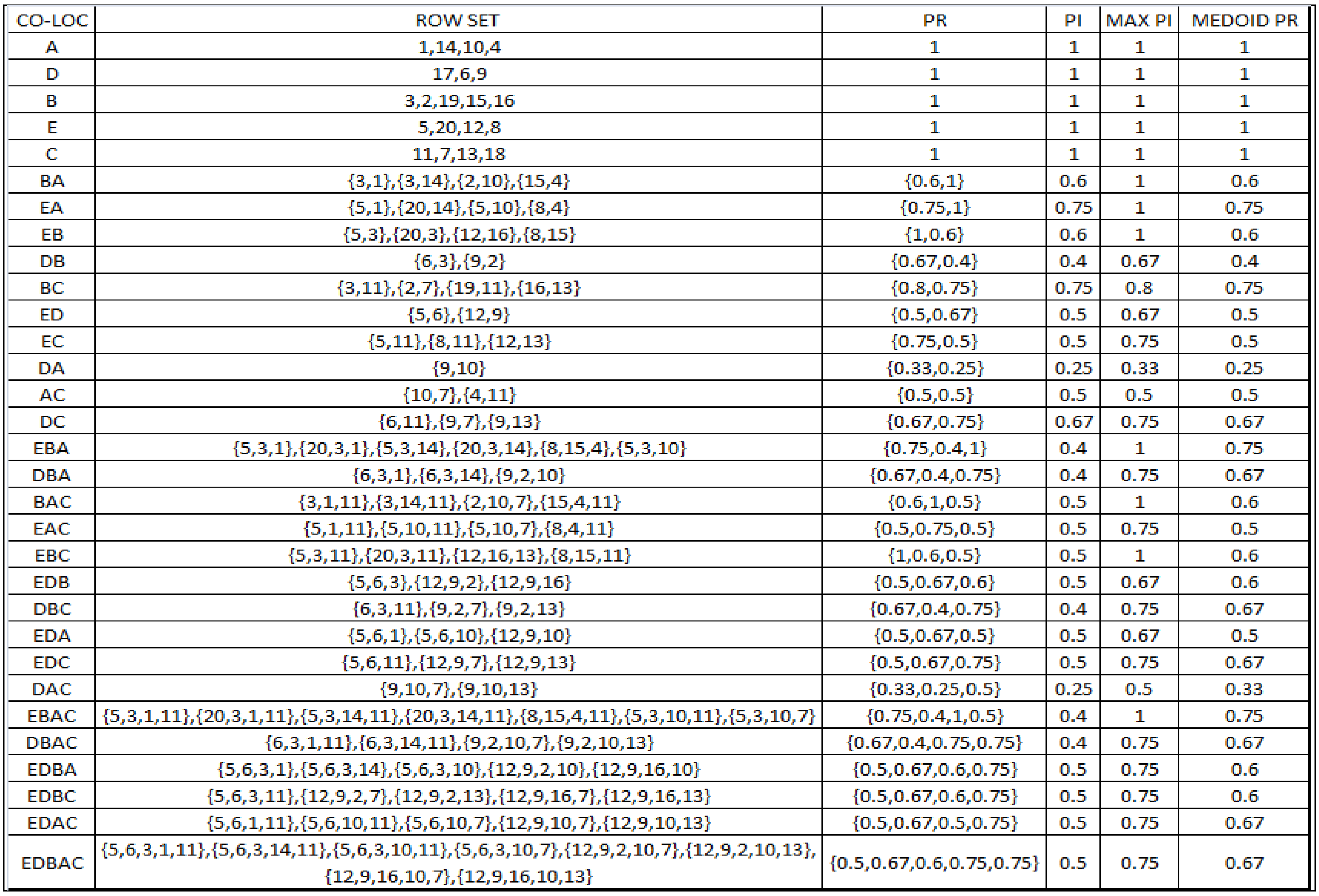


Figure 4: Plotting of co-location patterns



CO-LOC	ROW SET	PR	PI	MAX PI	MEDOID PR
A	1,14,10,4	1	1	1	1
D	17,6,9	1	1	1	1
B	3,2,19,15,16	1	1	1	1
E	5,20,12,8	1	1	1	1
C	11,7,13,18	1	1	1	1
BA	{3,1},{3,14},{2,10},{15,4}	{0.6,1}	0.6	1	0.6
EA	{5,1},{20,14},{5,10},{8,4}	{0.75,1}	0.75	1	0.75
EB	{5,3},{20,3},{12,16},{8,15}	{1,0.6}	0.6	1	0.6
DB	{6,3},{9,2}	{0.67,0.4}	0.4	0.67	0.4
BC	{3,11},{2,7},{19,11},{16,13}	{0.8,0.75}	0.75	0.8	0.75
ED	{5,6},{12,9}	{0.5,0.67}	0.5	0.67	0.5
EC	{5,11},{8,11},{12,13}	{0.75,0.5}	0.5	0.75	0.5
DA	{9,10}	{0.33,0.25}	0.25	0.33	0.25
AC	{10,7},{4,11}	{0.5,0.5}	0.5	0.5	0.5
DC	{6,11},{9,7},{9,13}	{0.67,0.75}	0.67	0.75	0.67
EBA	{5,3,1},{20,3,1},{5,3,14},{20,3,14},{8,15,4},{5,3,10}	{0.75,0.4,1}	0.4	1	0.75
DBA	{6,3,1},{6,3,14},{9,2,10}	{0.67,0.4,0.75}	0.4	0.75	0.67
BAC	{3,1,11},{3,14,11},{2,10,7},{15,4,11}	{0.6,1,0.5}	0.5	1	0.6
EAC	{5,1,11},{5,10,11},{5,10,7},{8,4,11}	{0.5,0.75,0.5}	0.5	0.75	0.5
EBC	{5,3,11},{20,3,11},{12,16,13},{8,15,11}	{1,0.6,0.5}	0.5	1	0.6
EDB	{5,6,3},{12,9,2},{12,9,16}	{0.5,0.67,0.6}	0.5	0.67	0.6
DBC	{6,3,11},{9,2,7},{9,2,13}	{0.67,0.4,0.75}	0.4	0.75	0.67
EDA	{5,6,1},{5,6,10},{12,9,10}	{0.5,0.67,0.5}	0.5	0.67	0.5
EDC	{5,6,11},{12,9,7},{12,9,13}	{0.5,0.67,0.75}	0.5	0.75	0.67
DAC	{9,10,7},{9,10,13}	{0.33,0.25,0.5}	0.25	0.5	0.33
EBAC	{5,3,1,11},{20,3,1,11},{5,3,14,11},{20,3,14,11},{8,15,4,11},{5,3,10,11},{5,3,10,7}	{0.75,0.4,1,0.5}	0.4	1	0.75
DBAC	{6,3,1,11},{6,3,14,11},{9,2,10,7},{9,2,10,13}	{0.67,0.4,0.75,0.75}	0.4	0.75	0.67
EDBA	{5,6,3,1},{5,6,3,14},{5,6,3,10},{12,9,2,10},{12,9,16,10}	{0.5,0.67,0.6,0.75}	0.5	0.75	0.6
EDBC	{5,6,3,11},{12,9,2,7},{12,9,2,13},{12,9,16,7},{12,9,16,13}	{0.5,0.67,0.6,0.75}	0.5	0.75	0.6
EDAC	{5,6,1,11},{5,6,10,11},{5,6,10,7},{12,9,10,7},{12,9,10,13}	{0.5,0.67,0.5,0.75}	0.5	0.75	0.67
EDBAC	{5,6,3,1,11},{5,6,3,14,11},{5,6,3,10,11},{5,6,3,10,7},{12,9,2,10,7},{12,9,2,10,13},{12,9,16,10,7},{12,9,16,10,13}	{0.5,0.67,0.6,0.75,0.75}	0.5	0.75	0.67

Figure 3: Co-location rules

5. CONCLUSION

In this paper co-location patterns are extracted using the prevalence measures using participation index approach. The co-location patterns are based on both the participation index and the *MaxPI.* The solution in this paper is based on the spatial dataset given in figure 1. Co-location patterns are extracted for all the five spatial objects in the dataset as shown in figure 2. *MaxPI* is used to get the most crucial objects among the five involved in the event of road accident. Hence, the solution has proposed a result which gives the occurrence of an event due to some highly influenced spatial objects. Also, where MaxPI concept fails, a new concept of MedoidPI has been used to get a proper solution to the problem. In future, a solution can be proposed to work on very large spatial databases and these databases can be real-time.

6. REFERENCES

[1] Neelammadhab Pandhy, Dr. Pragnyaban Mishra, Rasmita Panigrahi, "The Survey of Data Mining Applications and Feature Scope", IJCSEIT, Vol.2, No.3, June-2012.

[2] Shashi Shekhar, Siva Revada, Xuan Liu, "Spatial Databases – Accomplishments and Reasearch Needs", IEE transactions on knowledge and Data Engineering, Vol.11, 1999.

[9]

[3] Diansheng Guo, Jeremy Mennis, "Spatial Data Mininig and Geographic Knowledge Discovery- An Introduction", Computers, Environment and Urban Systems 33 (2009) 403-408.

[4] Yan Huang, Shashi Shekhar, Hui Xiong, " Discovering Co-location Patterns from Spatial Datasets: A General Approach", IEEE Transactions on Knowledge and Data Engineering.

[5] G.Kiran Kumar, P.Premchand, T. Venu Gopal, " Mining of Co-location Pattern from Spatial Datasets", IJCA (0975-8887), Vol.42 – No.21, March 2012.

[6] Yan Huang, Jian Pei, Hui Xiong, " Mining of Co-location Patterns with Rare Events from Spatial Datasets", Geoinformatica (2006) 10: 239-260.

[7] Richard Frank, Martin Ester, Arino Knobbe, " A Multi-Relational Approach to Spatial Classification", SIGKDD'09, Jun 28-July 1, 2009, Paris, France.

[8] Raymond T. Ng, Jiawei Han, " Efficient and Effective Clustering Methods for Spatial Data Mining", VLDB Conference Santiago, Chile, 1994.