International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 12 (2015) pp. 31269-31280

http://www.ripublication.com

Co-location Data Mining on Uncertain Datasets Using a Probabilistic Approach

M.Sheshikala[1] , D. Rajeswara Rao[2], and Md. Ali Kadampur[3]

[1]*SR Engineering College, Telangana, marthakala08@gmail.com*
[2]*KL University, Andhra Pradesh, rajeshduvvada@kluniversity.in*
[3]*SR Engineering College, Telangana, ali.kadampur@gmail.com*

Abstract

Uncertain data sets generally contain the real world data, such as mobile data, crime data, GIS data etc.,. Handling such data is a challenge for knowledge discovery particularly in colocation mining. Finding Probabilistic Prevalent colocations (PPCs) is one of the straight forward approach.. This method tries to find all colocations that are to be generated from a random world. For this we first apply an approximation error to find all the PPCs which reduce the computations. Next find all the possible worlds and split them into two different worlds and compute the prevalence probability. These worlds are used to compare with a minimum probability threshold to decide whether it is Probabilistic Prevalent colocation (PPCs) or not. The experimental results on the selected data set show the significant improvement in computational time in comparison to some of the existing methods used in colocation mining.

Index Terms— Probabilistic Approach, Colocation Mining, Un-certain Data Sets

I. INTRODUCTION

Basically colocation mining is the sub-domain of data mining. The research in colocation mining has advanced in the recent past addressing the issues with applications, utility and methods of knowledge discovery. Many techniques inspired by data base methods (Join based, Join-less, Space Partitioning, etc.,) have been attempted to find the prevalent colocation patterns in spatial data. Fusion and fuzzy based methods have been in use. However due to growing size of the data and computational time requirements highly scalable and computationally time efficient

framework for colocation mining is still desired. This paper presents a computational time efficient algorithm based on Probabilistic approach in the uncertain data.

Consider a spatial data set collected from a geographic space which consists of features like birds (of different types), rocks, different kinds of trees, houses, which is shown in Fig: 4. From this the frequent patterns on a spatial dimension can be identified, for example, < *bird, house* > and < *tree, rocks*>, the patterns are said to be colocated and they help infer a specific eco-system. This paper presents a computationally efficient method to identify such prevalent patterns from spatial data sets.

Since the object data is scattered in space (spatial coordinates) extracting information from it is quite difficult due to complexity of spatial features, spatial data types, and spatial relationships.

For example, a cable service provider may be interested in services frequently requested by geographical neighbours, and thus gain sales promotion data. The subscriber of the channel is located on a wide geographical positions and has wide ranging interest/preferences. Further in the process of collecting data there may be some missing links giving rise to uncertainty in the data. From the data mining point of view all this adds to complexity of analysis and needs to be handled properly. The paper addresses the uncertainty and data complexity issues in finding prevalent colocations.

The paper includes 1.The methods for finding the exact Probabilistic Prevalent colocations (PPCs). 2. Developing a dynamic programming algorithm to find Probabilistic Prevalent colocations (PPCs) which dramatically reduces the computation time. 3. Results of application of the proposed method on different data sets.

The remaining paper is organized as follows: In Section-1, we discuss the introduction, and related work is discussed in Section-2. In section-3 we discuss the definitions, and a block diagram to show the complete flow to find PPCs are discussed in section-4, In section-5 we discuss dynamic- programming algorithm for finding all Probabilistic Prevalent Colocations. We show the experiment results in Section-6. Finally, in section-7 we suggest future work.

II. RELATED WORK

Many methods have been extensively explored in order to find the Prevalent colocations in spatially Precise data. Some of these methods are:

Space Partitioning Method helps in finding the nearest objects of a subset of features. This approach may generate incorrect colocation patterns, because it may miss some of the colocation instances across partition.

Join Based approach finds the correct and complete colocation instances, This approach is computationally expensive with the increase of colocation patterns and their instances. Join-Less Approach: The join-less approach allocates the neighbor relationship between instances into a compressed star neighborhood. but the computation time of generating colocation table instances will increase with the growing length of colocation pattern. The author Yoo etal.[9],[10] has discussed the 2

algorithms , one among these is partial-join algorithm and the other is join-less algorithm. but in this approach there are some repeated scanning of materialized neighborhoods. Wang et al. [11]: A CPI-tree-based approach was developed by storing star-neighborhoods in a more compact format and a prefix tree instead of a table, which reduces the repeated scans of materialized neighborhoods as in[9]. In this paper [12] discovered colocation patterns from interval data. As different applications are growing the researchers are more devoted to extend the traditional frequent pattern mining to uncertain data sets. [1], [2], [3]. Chui et al.[3] : Proposed a method which accurately mine the frequent patterns maintaining the efficiency, later in paper [4], methods were used for finding the frequent items in very large uncertain data sets Besides the above representative colocation mining problem, in this paper we are closely related to finding the prevalent colocations using the Probabilistic approximation approach[13]. Huang et al.[6] In this paper a general framework was proposed for a prior-gen based colocation mining, in which minimum-participation ratio measure was taken instead of support, in which anti-monotone property which increases the computational efficiency. Later a paper[14],[16] was published which proposed a join-based algorithm to find prevalent colocation patterns, but as the size of the data set grows the number of joins increases. Later Huang et al. extended the problem to mining confident colocation patterns in which maximum participation ratio was taken instead of minimum participation ratio which is used to measure the prevalence of confident colocation

III. THE BASIC DEFINITIONS

A. Uncertain Data Sets:

Uncertain data set is defined as the data that may contain errors or may only be partially complete. Many advanced technologies have been developed to store and record large quantities of data continuously. In many cases, For example there are different types of features like tree, Bird, Rocks and House and we have instances for the features like trees which are of various types of trees, and Birds which are like Eagle, Sparrow, Owl, and the Features like rock and house are having only one kind of instance. From the figure we can conclude that rocks and a type of tree is colocated, Sparrow and house are colocated.

We can identify that there are different types of features like tree, Bird, Rocks and House and we have instances for the features like trees which are of various types of trees, and Birds which are like Eagle, Sparrow, Owl, and the Features like rock and house are having only one kind of instance. From the figure we can conclude that rocks and a type of tree is colocated, Sparrow and house are co-located.

B. Instance of a Feature:

The instances of a feature are the existential probability of the instance in the place location. If F is a feature then F.i is an instance.

TABLE I A SAMPLE EXAMPLE OF SPATIAL UNCERTAIN DATA SET

Id if Instance w	Spatial Feature	Location	Probability
1	A1	in Fig.1	0.1
2	A2	in Fig.1	0.4
3	A3	in Fig.1	0.7
4	B1	in Fig.1	0.1
5	B2	in Fig.1	1
6	C1	in Fig.1	1
7	C2	in Fig.1	0.1
8	D1	in Fig.1	0.4
9	D2	in Fig.1	0.1

*C. **Spatially Uncertain Feature:***

A spatial feature contains the spatial instances, and a data set Z containing spatially uncertain features is called spatially uncertain data set. If Z is a data set then set of features is A, B, C,...

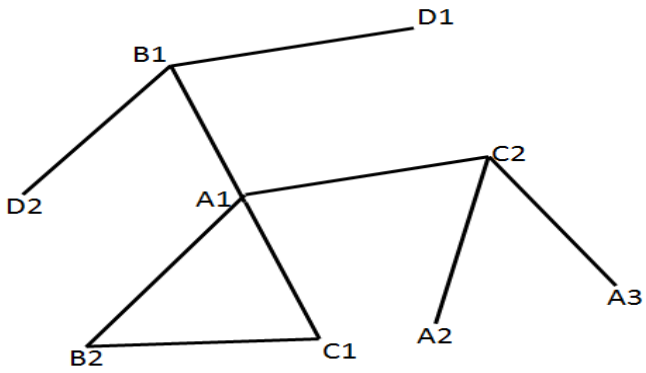

Fig: 1 Distribution of example spatial Instance

*D. **Probability of Possible Worlds***

For each colocation of k-size, c={$f_1, f_2, \ldots\ldots\ldots f_n$}of each instance $F.i$ there are two different possible worlds (i) one among them is that the instance is present (ii) and the other is absent. Take the set of features F={$f_1, f_2, \ldots\ldots\ldots f_n$} and the set of instances S= {$S_{f_1}, S_{f_2}, \ldots\ldots\ldots, S_{f_n}$} , where S_{f_i} $(1 <= i <= k)$ is the set of instances in S and there are $2^{|S|}=2^{|s_{f_1} s_{f_2} \ldots\ldots\ldots s_{f_n}|}$ possible worlds at most. Each Possible world w is associated with a probability P (w) that is the true world, where P (w) > 0.

E. ***Neib_tree***

The Neib_tree is constructed for the Table-I which indicates the existence of the path from one feature to the other. If there is a path it indicates that a table instance is existing. This Neighbouring tree eliminates the duplicates can be seen in Fig:2.

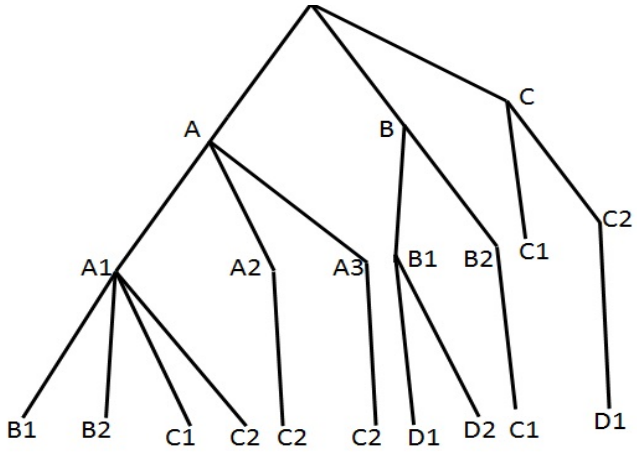

Fig:2 Neib_tree for Fig:1

TABLE II COMPUTATIONAL PROCESS OF COLOCATION (A,C)

A1	A2	A3	C1	C2
0	0	0	0	0
0	0	0	0	1
0	0	0	1	0
0	0	0	1	1
0	0	1	0	0
0	0	1	0	1
0	0	1	1	0
0	0	1	1	1
0	1	0	0	0
0	1	0	0	1
0	1	0	1	0
0	1	0	1	1
0	1	1	0	0
0	1	1	0	1
0	1	1	1	0
0	1	1	1	1
1	0	0	0	0
1	0	0	0	1
1	0	0	1	0
1	0	0	1	1

1	0	1	0	0
1	0	1	0	1
1	0	1	1	0
1	0	1	1	1
1	1	0	0	0
1	1	0	0	1
1	1	0	1	0
1	1	0	1	1
1	1	1	0	0
1	1	1	0	1
1	1	1	1	0
1	1	1	1	1

TABLE III COMPUTATIONAL PROCESS OF COLOCATION(A,C)

Possible World$_w$	P(w_i)
w_1={C1}	0:1458
w_2={C1,C2}	0:0162
w_3={A3,C1}	0:3402
w_4={A3,C1,C2}	0:0378
w_5={A2,C1}	0:0972
w_6={A2,C1,C2}	0:0108
w_7={A2,A3,C1}	0:2268
w_8={A2,A3,C1,C2}	0:0252
w_9={A1,C1}	0:0162
w_{10}={A1,C1,C2}	0:0018
w_{11}={A1,A3,C1}	0:0378
w_{12}={A1,A3,C1,C2}	0:0042
w_{13}={A1,A2,C1}	0:0108
w_{14}={A1,A2,C1,C2}	0:0012
w_{15}={A1,A2,A3,C1}	0:0252
w_{16}={A1,A2,A3,C1,C2}	0:0028

IV. BLOCK DIAGRAM

Basic flow of co-location pattern mining: In this section, we present a flow diagram which describes the flow of identifying the Probabilistic Prevalent colocations. Given a Spatial data set, a neighbor relationship, and interest measure thresholds the basic colocation pattern mining involves 4 steps as in Fig:3

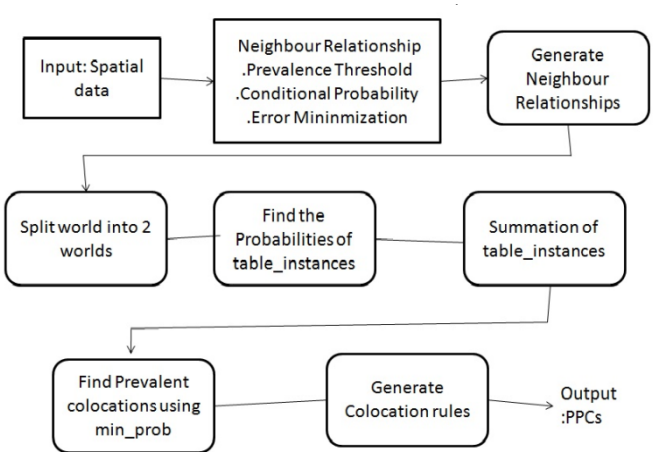

Fig:3 Block diagram to find the PPCs

First candidate colocation patterns are generated and the colocation instances and spitted into two worlds from the spatial data set. Next, find the probabilities using minimum prevalence and compute summation of table instances of each colocation, Next find prevalent colocation using minimum probability.

The Implementation of the Algorithm:

1. Reading the value of e:

if the value of e is 1 then the algorithm stops and prints that all colocations are Prevalent. Otherwise if the value is in between 0<e<1 then execute steps from 4 to 14.

2. The Initializing Steps:

After finding all neighboring instance pairs, a Neib_tree can be generated using the method [5]. For example Fig:2 are a Neib_tree generate from Fig:1 These Neib_tree consist of a set of features which are organized in ordered and branched form.

2. 1: Generating Coarse Combination instances from each colocation:

This step computes the coarse combinations of different colocation of k-size. For example for colocation (A,C) we get a set of 24 combination instances out of 25 combinations whose probability is greater than zero.

2. 2: Splitting of Colocation instances:

Splitting of a colocation into two different worlds (i.e.)., colocation based on the set of features which has largest number of instances. W_1 is the set of possible worlds of $\{f_1\}$ and W_2 is that of possible set of worlds of $\{f_1, f_2, \ldots\ldots\ldots, f_n\}$. For example in this paper the Colocation (A, C) are divided into 2 worlds out of which W_1 in consisting of all instances of {A} & {A, C} and W_2 consisting alone {C} instances (i.e.), {C1} & {C1, C2}.

2. 3: Computing the Probability of table instances in world W_2:
Computing the Probability of table instances W_2 where W_2 is consisting of ({C1},{C1,C2}). (i,e)., *for* $Pr(c, f_1)[i, j]\{C2\}$ and $Pr(c, \bar{f}_1)[i, j]\{C2\}$. After finding the Probabilities the values can be seen in TABLE IV and V :

TABLE IV THE COMPUTATION OF THE $P(c, f_1)[i, j]\{C1\}$ AND $P(c, \bar{f}_1)[i, j]\{C1\}$

	j =0	j =1	j =2	j =3
i=0	(1,1)	(0.7,1)	(0.7,0.6)	(0.7,0.24)
i =1	(0,0)	(0.3,0)	(0.3,0.4)	(0.3,0.52)
i =2	(0,0)	(0,0)	(0,0)	(0,0.24)
i =3	(0,0)	(0,0)	(0,0)	(0,0)

TABLE V THE COMPUTATION OF THE $P(c, f_1)[i, j]\{C1, C2\}$ AND $P(c, \bar{f}_1)[i, j]\{C1, C2\}$

	j=0	j =1	j =3	j =4
i =0	(1,1)	(0.7,1)	(0.42,1)	(0.42,0.4)
i =1	(0,0)	(0.3,0)	(0.46,0)	(0.46,0.6)
i =2	(0,0)	(0,0)	(0.12,0)	(0.12,0)
i =4	(0,0)	(0,0)	(0,0)	(0,0)

A. ***Computing the Prevalence Probability of world W_2:***
After computing step-9 for each set colocation of k-size, now compute the prevalence probability from equation: 5 For example for colocation (A,C) if the min_prev is 0.5 then for the table_ instances { C1} the value is 0.205 and for{ C1, C2 } the value is 0.058.

B. ***Summation of Prevalence Probabilities:***
After computing the prevalence Probability of all colocation then we make the summation of all Prevalence Probability. For example for colocation (A, C) the value of {C1} is 0.2052 and {C1, C2} is 0.058 and the summation of both { C1 } & {C1, C2} is 0.2632

C. ***: Checking with Minimum Probability:***
if the summation is less than the minimum probability then it is removed from Probabilistic Prevalent Colocations, Otherwise added to prevalent Colocation. From the above example if the min_prob is 0.3 then colocation (A, C) is filtered and if it is 0.2 then colocation is selected.

3. The colocation size is increased and next steps are executed.
4. Once all the Probabilistic Prevalent Colocations are identified the algorithm stops.
5. A Union of all Probabilistic Prevalent Colocations is written from a set of features.

V. RESULTS

The results are compared against a data set given in the following Table-VI which consists of 7 features with an average of 2 instances.

TABLE VI A SYNTHETIC SAMPLE DATA SET

Features	X-Coordinates	Y-Coordinates	Probability
0	328	1362	0.5
0	190	1140	0.4
0	392	1220	0.9
1	290	1264	0.1
1	330	1480	1
2	260	1278	0.1
3	185	1440	0.1
3	320	1500	0.4
3	330	1500	0.7
4	150	1580	0.1
4	150	1300	1
5	225	1300	1
5	260	1530	0.1
6	220	1650	0.4
6	60	1590	1

From Table-VI we get 2 PPCs when $min\_prev = 0.4$ and $min\_prob = 0.4$ and d=150, and = 0.001 and those PPCs are {1, 3} and {4, 5},the result can be seen in the following Fig:4

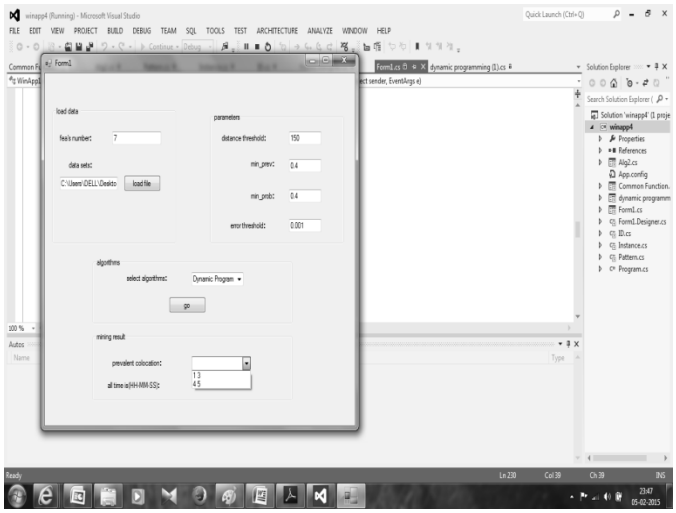

Fig:4 PPCs for Table-VI with $min\_prev = 0.4$ and $min\_prob = 0.4$, d=150,and $\varepsilon = 0.001$

Likewise when the comparisons are made against the complete data set from Table-VI we get the following Prevalent and non-Prevalent colocations, varying the *min_prev* and *min_prob* for the distance threshold=150 which are shown in Fig:6.

<min_prev,min_prob>	Prevalent Colocations	Non Prevalent Colocations
<0.2,0.2>	(0,1)(0,3)(0,5)(1,3)(4,5)(0,1,3)	(0,2)(0,4)(1,2)(1,4)(1,5)(2,3)(2,4)(2,5)(3,4)(3,5)(0,1,2)
<0.2,0.4>	(0,1)(0,3)(0,5)(1,3)(4,5)(0,1,3)	(0,2)(0,4)(1,2)(1,4)(1,5)(2,3)(2,4)(2,5)(3,4)(3,5)(0,1,2)
<0.2,0.6>	(1,3)(4,5)	(0,2)(0,3)(0,4)(0,5)(1,2)(1,4)(1,5)(2,3)(2,4)(2,5)(3,4)(3,5)(0,1,2)(0,1,3)
<0.2,0.8>	(1,3)(4,5)	(0,2)(0,3)(0,4)(0,5)(1,2)(1,4)(1,5)(2,3)(2,4)(2,5)(3,4)(3,5)(0,1,2)(0,1,3)
<0.4,0.2>	(0,1)(0,3)(0,5)(1,3)(4,5)(0,1,3)	(0,2)(0,4)(1,2)(1,4)(1,5)(2,3)(2,4)(2,5)(3,4)(3,5)(0,1,2)
<0.4,0.4>	(1,3)(4,5)	(0,2)(0,3)(0,4)(0,5)(1,2)(1,4)(1,5)(2,3)(2,4)(2,5)(3,4)(3,5)(0,1,2)(0,1,3)
<0.4,0.6>	(1,3)(4,5)	(0,2)(0,3)(0,4)(0,5)(1,2)(1,4)(1,5)(2,3)(2,4)(2,5)(3,4)(3,5)(0,1,2)(0,1,3)
<0.4,0.8>	(1,3)(4,5)	(0,2)(0,3)(0,4)(0,5)(1,2)(1,4)(1,5)(2,3)(2,4)(2,5)(3,4)(3,5)(0,1,2)(0,1,3)
<0.6,0.2>	(1,3)(4,5)	(0,2)(0,3)(0,4)(0,5)(1,2)(1,4)(1,5)(2,3)(2,4)(2,5)(3,4)(3,5)(0,1,2)(0,1,3)
<0.6,0.4>	(1,3)(4,5)	(0,2)(0,3)(0,4)(0,5)(1,2)(1,4)(1,5)(2,3)(2,4)(2,5)(3,4)(3,5)(0,1,2)(0,1,3)
<0.6,0.6>	(1,3)(4,5)	(0,2)(0,3)(0,4)(0,5)(1,2)(1,4)(1,5)(2,3)(2,4)(2,5)(3,4)(3,5)(0,1,2)(0,1,3)
<0.6,0.8>	(4,5)	(0,2)(0,3)(0,4)(0,5)(1,2)(1,3)(1,4)(1,5)(2,3)(2,4)(2,5)(3,4)(3,5)(0,1,2)(0,1,3)

Fig:6 PPCs for Table-VI with varying *min_prev* and *min_prob*, and d=150,and ε = 0.001

From the graph below it is proved that the computation time for the improved Approximation algorithm works well when compared to dynamic algorithm: as shown in Fig:6

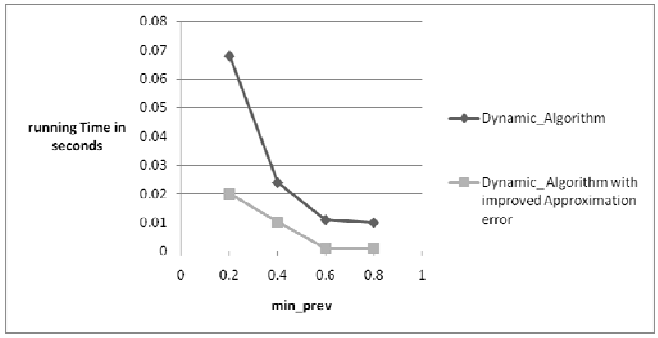

Fig:6 Varying *min_prev* and *min_prob*, d=150, and ε= 0.001

VI. CONCLUSION

The Proposed method is for finding Probabilistic Prevalent Colocation in Spatially Uncertain data sets which are likely to be prevalent. We have given an approach in which the computation time is drastically reduced. Further keeping in view the work can be extended to find the important sub functionalities in colocation mining to formulate colocation mining specific primitives for the next generation programmer

which we can expect to evolve as a scripting language. In essence the scope of the work can cover data base technologies, parallel programming domain, graphical graph methods, programming language paradigms and software architectures.

REFERENCES

[1] C.C. Aggarwal et al, ”Frequent Pattern Mining with Uncertain Data,” Proc. 15th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining(KDD), pp. 29-37, 2009.

[2] T. Bernecker, H-P Kriegel, M. Renz, F. Verhein, and A. Zuefle, ”Proba-bilistic Frequent Itemset Mining in Uncertain Databases,” Proc. 15th ACM SIGKDD Conf. Knowledge Discovery and Data Mining(KDD ‘09), pp. 119-127, 2009. [05]

[3] C.-K. Chui, B. Kao, and E. Hung, ”Mining Frequent Item sets from Uncertain Data,” Proc. 11th Pacific-Asia Conf. Knowledge Discovery and Data Mining(PAKDD), pp. 47-58, 2007.

[4] C.-K. Chui, B. Kao, ”A Decremental Approach for Mining Frequent Item sets from Uncertain Data,” Proc. 12th Pacific-Asia Conf. Knowledge Discovery and Data Mining(PAKDD), pp. 64-75, 2008.

[5] Y. Huang, H. Xiong, and S. Shekar, ”Mining Confident Co-Location Rules Without a Support Threshold,” Proc. ACM Symp. Applied Com-puting, pp. 497-501, 2003.

[6] Y. Huang, S. Shekar, and H. Xiong, ”Discovering Co-Location Patterns from Spatial Data Sets: A General Approach,” IEEE Trans. knowledge and Data Eng., vol. 16, no. 12, pp. 1472-1485, Dec. 2004.

[7] Y. Huang, J. Pei, and H. Xiong, ”Mining Co-Location Patterns with Rare Events from Spatial Data Sets,” Geoinformatics, vol. 10, no. 3, pp. 239-260, Dec. 2006.

[8] Y. Morimoto, ”Mining Frequent Neighbouring Class Sets in Spatial Databases,” Proc. Seventh ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining(KDD), pp. 353-358, 2001.

[9] J.S. Yoo, S. Shekar,J. Smith, and J.P. Kumquat, ”A Partial Join Approach for Mining Co-Location Patterns,” Proc. 12th Ann. ACM Int’l Workshop Geographic Information Systems (GIS), pp. 241-249, 2004.

[10] J.S. Yoo and S. Shekar, ”A Join less Approach for Mining Spatial Co-Location Patterns,” IEEE Trans. knowledge and Data Eng.(TKDE), vol. 18, no. 10, pp. 1323-1337, Dec. 2006.

[11] L. Wang, Y. Bao, J. Lu and J. Yip, ”A New Join-less Approach for Co-Location Pattern Mining,” Proc. IEEE Eighth ACM Int’l Conf. Computer and Information Technology (CIT), pp. 197-202, 2008.

[12] L. Wang, H. Chen, L. Zhao and L. Zhou, ”Efficiently Mining Co-Location Rules of Interval Data,” Proc. Sixth Int’l Conf. Advanced Data Mining and Applications, pp. 477-488, 2010.

[13] Q. Zhang, F. Li, and K. Yi, ”Finding Frequent Items in Probabilistic Data,” Proc. ACM SIGMOD Int’l Conf. Management of Data, pp. 819-832, 2008.

[14] L. Wang, P. Wu, and H. Chen, ”Finding Probabilistic Prevalent Colocations in Spatially Uncertain Data Sets,” IEEE Trans. knowledge and Data Eng.(TKDE), vol. 25, no. 4, pp. 790-804, Apr. 2013.