MINING CO-LOCATION PATTERNS WITH CLUSTERING ITEMS FROM SPATIAL DATA SETS

Guoqing Zhou [1], Qi Li [1, 2], Guangming Deng [2, *], Tao Yue [1], Xiang Zhou [1]

[1] Guangxi Key Laboratory of Spatial Information and Geomatics, Guilin University of Technology, No. 12 Jian'gan Road, Guilin, Guangxi 541004, China - (gzhou, yuetao, zqx0711)@glut.edu.cn
[2] College of Science, Guilin University of Technology, No. 12 Jian'gan Road, Guilin, Guangxi 541004, China - dgm@glut.edu.cn

Commission Ⅲ, WG Ⅲ/1

KEY WORDS: Spatial data mining, Clustering items, Co-location patterns with clustering items, Neighbor relationship, Praticipation index

ABSTRACT:

The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the spatial data mining. Co-location patterns discovery is an important branch in spatial data mining. Spatial co-locations represent the subsets of features which are frequently located together in geographic space. However, the appearance of a spatial feature C is often not determined by a single spatial feature A or B but by the two spatial features A and B, that is to say where A and B appear together, C often appears. We note that this co-location pattern is different from the traditional co-location pattern. Thus, this paper presents a new concept called clustering terms, and this co-location pattern is called co-location patterns with clustering items. And the traditional algorithm cannot mine this co-location pattern, so we introduce the related concept in detail and propose a novel algorithm. This algorithm is extended by join-based approach proposed by Huang. Finally, we evaluate the performance of this algorithm.

1. INTRODUCTION

As an important research direction in the field of spatial data mining, Co-location patterns discovery has a wide range of applications, include ecology, Earth science, biology, public health, transportation, etc. Similarly, the co-location patterns with clustering items are also of great significance to the above fields of application. Algorithms of traditional co-location mining cannot be used for mining co-location patterns with clustering items directly. Therefore, we conduct a detailed study of the novel co-location patterns and present an algorithm for mining it.

1.1 Related Works

In previous works on co-location patterns discovery，the concept of co-location patterns with clustering items has not been discussed. For traditional co-location patterns, the previous literature proposed different mining algorithms. Huang, Shekhar and Xiong (2004) proposed a general approach: Join-based approach. At the same time, they defined participation index that has an anti-monotone property. Furthermore, they showed the relationship between the participation index and a spatial statistics interest measure, the cross-K function. Yoo and Shekhar developed the partial-join (2004) and the joinless (2005) approaches to mining co-location patterns, the two algorithms greatly reduce the computational cost. Huang, Pei and Xiong (2006) addressed the problem of mining co-location patterns with rare spatial events. In their paper, a new measure called the maximal participation ratio (maxPR) was introduced and a weak monotonicity property of the maxPR measure was identified. Xiao et al (2008) introduced the density based co-location pattern discovery. The concept of the negative co-location patterns was defined by Jiang et al (2010). Based on the analysis of the relationship between negative and positive participation index, they proposed methods for negative participation index calculation and negative patterns pruning strategies. Zhou et al (2012) applied co-location patterns to the decision tree, they developed a called co-location decision tree (CL-DT) method.

1.2 Our Contributions

In this paper, the definition of clustering items is given, and we present a novel co-location pattern, i.e. co-location patterns with clustering items. First the basic concepts of co-location patterns with clustering items and rules are defined. Second, we study the problem of efficiently mining co-location patterns with clustering items systematically. Through the review of the previous approaches, we propose a novel approach for mining co-location patterns with clustering items based on the join-based approach. Finally, we conduct experimental evaluation use a synthetic dataset. The results show that our algorithm is correct and efficient.

2. BASIC CONCEPTS

Definition 1（clustering items） $X = \{f_i, f_j\}$ is a clustering item if f_i and f_j satisfy the neighbor relationship. We also write this clustering item as $f_i f_j$.

Example 1 In Fig.1, A.2B.6 is a clustering item, if A.2 and B.6 are neighbor i.e. $\mathrm{distance}(A.2, B.6) \leq d$.

* Corresponding author: Guangming Deng; E-mail: dgm@glut.edu.cn

Definition 2 （spatial neighbor relationship） If R is defined as a Euclidean distance metric and its threshold value is d,

(1) two spatial objects are neighbors if they satisfy the spatial neighbor relationship: $R(f_i, f_j) \leftrightarrow (distance(f_i, f_j) \leq d)$;

(2) a spatial object and a clustering item are neighbors if they satisfy the spatial neighbor relationship:

$\mathrm{R}(f_k, f_i f_j) \leftrightarrow (distance(f_k, f_i) \leq d, distance(f_k, f_j) \leq d)$.

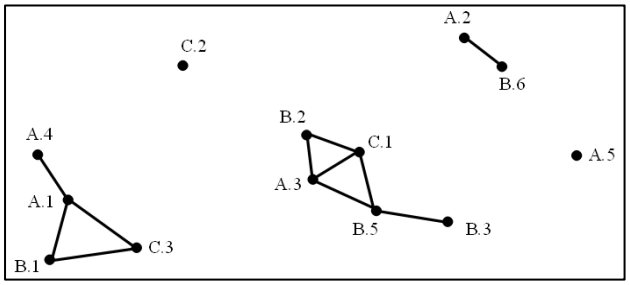

Figure 1. An example data set

Example 2 In Figure 1, C.3 and clustering item A.1B.1 satisfy the neighbor relationship, because C.3 and A.1,B.1, respectively, to satisfy the neighbor relationship, and A.1B.1 is a clustering item.

Definition 3 (row instance) T is a co-location pattern with clustering items, a neighborhood instance I of T is a row instance of T if I contains instance of all events in T and no proper subsets of I does so. The **table instance** of T is the collection of all row instance of T.

Definition 4(co-location patterns with clustering items ） T is a co-location pattern with clustering items, if $\mathrm{T} = \mathrm{X} \cup \mathrm{Y}$,where X is a clustering item,Y is a set of spatial features,$|X| = 2, |Y| \geq 1, X \cap Y = \emptyset$.

Definition 5 (the PR of co-location patterns with clustering items) The participation ratio $\mathrm{PR}(T, f_i)$ in a co-location pattern with clustering items $\mathrm{T} = \{f_i, \cdots, f_k\}$ is a fraction of feature f_i which participate in any row instance of co-location pattern with clustering items T.

$$\mathrm{PR}(T, f_i) = \frac{\left|\pi_{f_i}(table\_instance(T))\right|}{|table\_instance(f_i)|} \quad (1)$$

Where π is the relational projection operation with duplication elimination.

Example 3 In Fig.1,X={A,B}(it can also be written as AB) has four instances, i.e. A.1B.1， A.2B.6， A.3B.2 and A.3B.5.For a co-location pattern with clustering items T={X,C}(i.e. T={AB,C})， it's instances are {A.1B.1,C.3},{A.3B.2,C.1} and {A.3B.5,C.1}. $\mathrm{PR}(T, X) = 3/4$, because only A.1B.1, A.3B.2 and A.3B.5 appear in T's instances.

Definition 6 (the PI of co-location patterns with clustering items) The participation index of a co-location pattern with clustering items $\mathrm{T} = \mathrm{X} \cup \mathrm{Y}$ is defined as $\mathrm{PI}(T) = min\{PR(T, X), PR(T, Y)\}$.

Example 4 In Fig.1, for the co-location with clustering items T={AB,C}, $\mathrm{PI}(T) = min\{PR(T, AB), PR(T, C)\} = 2/3$, because $\mathrm{PR}(T, AB) = 3/4$ and $\mathrm{PR}(T, C) = 2/3$.

Definition 7（prevalence co-location patterns with clustering items ） min_prev is a minimum prevalence threshold. A co-location pattern $\mathrm{T} = \mathrm{X} \cup \mathrm{Y}$ is a co-location pattern with clustering items ($\mathrm{X} = \{f_i, f_j\}$), if T meets the following conditions.

(1) $\mathrm{PI}(f_i \cup Y) < \min\_prev$, $\mathrm{PI}(f_j \cup Y) < \min\_prev$, i.e. $\{f_i, Y\}$ and $\{f_j, Y\}$ are not prevalent co-location patterns.

(2) $\mathrm{PI}(T) \geq \min\_prev$.

Example 5 In Fig.1,easy to find C has 3 instances, the clustering set AB has four instance: A.1B.1, A.2B.6, A.3B.2 and A.3B.5. The clustering set AB can be combined with C into a size 2 candidate co-location with clustering items T={AB,C},it has three instances {{A.1B.1,C.3},{A.3B.2,C.1},{A.3B.5,C.1}}. We can calculate it:$\mathrm{PR}(T, AB) = 3/4$, $\mathrm{PR}(T, C) = 2/3$. Therefore, $\mathrm{PI}(T) = min(PR(T, AB), PR(T, C)) = 2/3$. And easy to calculate $\mathrm{PI}(A, C) = 0.4$, $\mathrm{PI}(B, C) = 0.5$. If minimum prevalence threshold min_prev is set to 0.6, T is a prevalent co-location with clustering items since it meets the condition (1) and condition (2).

Definition 8 (conditional probability) The conditional probability $\mathrm{CP}(X \rightarrow Y)$ of rules of co-locations with clustering items $\mathrm{X} \rightarrow \mathrm{Y}$ is the fraction of instance X in the neighborhood of instances of Y, i.e.,

$$\mathrm{CP}(X \rightarrow Y) = \frac{\left|\pi_X(table\_instance(\{X \cup Y\}))\right|}{|table\_instance(\{X\})|} \quad (2)$$

Definition 9 (rules of co-locations with clustering items) $\mathrm{X} \rightarrow \mathrm{Y}$ is a rule co-locations with clustering items if $\mathrm{X} \cup \mathrm{Y}$ is a prevalent co-location pattern with clustering items and the conditional probability of $\mathrm{X} \rightarrow \mathrm{Y}$ is more than a conditional probability threshold (min_conf) defined by users.

3. OUR APPROACH

3.1 Review of Join-based Approach

Huang, Y., Shekhar, S. and Xiong, H. (2004) proposed an instance join-based co-location mining algorithm. First, after finding all neighbor pair objects(size 2 co-location instances) using a geometric method, the method finds the instances of size k (> 2) co-locations by joining the instances of its size k-1 subset co-locations where the first k-2 objects are common. Fig.2 (b) shows the procedure to generate the instances of co-location {A, B, C}. The instances of co-location {A, B} and the instances of co-location {A, C} are joined with the first objects, and then the neighbor relationships between the second objects are checked. This approach finds correct and complete co-location instance sets.

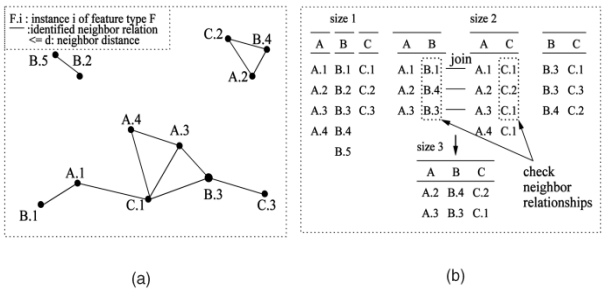

Figure 2. (a) Example dataset. (b) Instance join. (Yoo, J. S. and Shekhar, S., 2006)

3.2 Our Approach for Mining Co-location Patterns with Clustering Items

Algorithms of traditional co-location mining cannot be used for mining co-location patterns with clustering items directly, because some definitions of co-location patterns with clustering items had been redefined (such as the definition of spatial neighbor relationship between X and Y) and some methods must be redesigned (such as how to calculate the PI).

Our approach for mining co-location patterns with clustering items is extended by join-based approach,still using the principle of instance join. It has five phases. The first phase finds all clustering items. The second phase computes neighbor relationships between spatial instances and clustering items. The third phase generates size-k candidate co-locations with clustering items. The fourth phase is pruning. The fifth phase generates prevalent co-locations with clustering items and rules of co-locations with clustering items.

3.2.1 Discover All Clustering Items in Spatial Database

This is the basic step of this algorithm. As in Definition 1, we use Euclidean distance to measure whether two instances satisfy the neighbor relationship. Once two instances satisfy the neighbor relationship, we call them a clustering item. In this step we need to find out all the clustering items.

For co-locations with clustering items in this paper, $T = X \cup Y$ is a co-location of size k, if $|X| = 2, |Y| = k - 1$. And we note that co-locations with clustering items of size 1 are different from traditional co-locations of size 1. All the clustering items which we discovered are co-locations with clustering items of size 1. All co-locations with clustering items of size 1 are also prevalent and we need not calculate their prevalence measures, because the value of participation index is 1 for all co-locations with clustering items of size 1.

3.2.2 Generation of Candidate Co-locations

Similar to the traditional algorithm, we could rely on a combinatorial approach to generate size k+1 candidate co-locations with clustering items from size k prevalent co-locations. Specially, a clustering set and an instance are combined to generate a size 2 candidate co-location with clustering items.

3.2.3 Pruning

In traditional algorithm, Candidate co-locations can be pruned using the given threshold min _prev on the prevalence measure. The min _prev can also be used in our algorithm to prune candidate co-locations with clustering items. This kind of pruning method is called prevalence-based pruning by Huang[2]. For a candidate co-location with clustering items $T = X \cup Y(X = \{f_i, f_j\})$, not only do we have to calculate PI(T), but we also have to calculate PI(f_i, Y) and PI(f_j, Y). Because a prevalent co-location with clustering items must meet the condition (1) and condition (2) proposed in **Definition 7**.

In addition, we prevent a novel pruning method: due to condition (1), X which contains f_i cannot be combined with f_k into a prevalent co-location with clustering items, if $\{f_i, f_k\}$ is a prevalent co-location i.e. PI$(\{f_i, f_k\}) \geq \min_prev$.This method can greatly reduce the unnecessary calculation time.

For example, in Fig.1, {AC,B}, {BC,A} cannot be prevalent co-location with clustering items since {A,B} is a prevalent co-location. We can prune directly and there is not necessary to calculate PI$(\{AC, B\})$, PI$(\{BC, A\})$.

Input:	A spatial database S, a set of spatial feature types $F = \{f_1, f_2, \cdots, f_n\}$, a neighborhood relationship R, a minimum prevalent threshold min _prev, and a conditional probability threshold min _conf..
Output:	A set of co-locations with clustering items with prevalence and conditional probability values greater than user-specified minimum prevalence and conditional probability thresholds.
Variables:	k: co-location size Q: all clustering items in spatial database C_k: set of candidate size-k co-locations with clustering items P_k: set of prevalent size-k co-locations with clustering items in C_k
Method:	1. discover all clustering items in spatial database; 2. let k=2, generate C_2, the set of candidate 2-patterns: 3. for each $C \in C_k$ calculate PI(C), PI$(\{f_i, Y\})$ and PI$(\{f_j, Y\})$; ($C = \{X, Y\}$, f_i and f_j are different feature types in X) 4. let P_k be the subset of C_k such that for each $P \in P_k$, PI$(P) \geq \min_prev$, PI$(\{f_i, Y\}) \leq \min_prev$, PI$(\{f_j, Y\}) \leq \min_prev$; 5. generate the set C_{k+1} of candidate (k+1)-patterns; 6. if $C_{k+1} \neq \emptyset$, let k=k+1, go to step 2; 7.output $\cup_i P_i$.

Figure 3. The algorithm for mining co-location patterns with clustering items

4. EXPERIMENTAL EVALUATION

We evaluate this algorithm using synthetic datasets. Synthetic datasets were generated using a spatial data generator similar to the data generator used in Shekhar and Huang (2001). The number of spatial feature types is 20. Three parameters, namely number of spatial instances (n), prevalence threshold (min _prev), and spatial neighbor distance threshold (d), were varied during the experiments for verifying the effects of parameters and the performance of the algorithm.

4.1 Effect of Number of Spatial Instances

We examined the performance of the algorithm with the number of spatial instances. We used a spatial frame of 1000*1000. Once the number of spatial instances changes, the density of the data will change. As is shown in Fig.4, the execution time of this algorithm significantly increased with the increment of the number of spatial instances. This is very similar to the join-based algorithm, because as the number of spatial instances increases, a large number of joins are required.

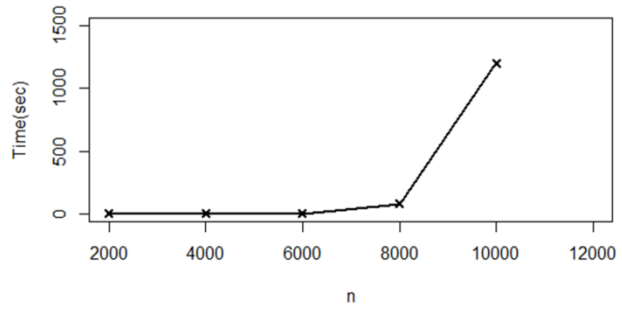

Figure 4. Effect of number of spatial instance

4.2 Effect of Parameter min_prev

The following experiment examined the effect of parameter min _prev for running time. In the experiment, the number of spatial instances is 10K, and the parameter d is set to 20. As is shown in Fig.5, when the prevalence threshold changes from 0.4 to 0.8, the execution time does not change much. However, when the prevalence threshold decreases from 0.4, the running time starts to increase rapidly.

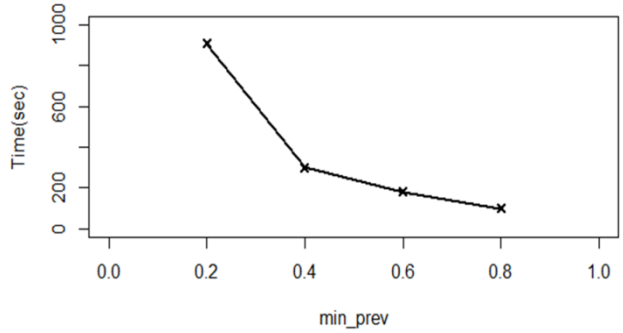

Figure 5. Effect of Parameter min_prev

4.3 Effect of Parameter d

This experiment examined the effect of parameter d for running time. In the experiment, the number of spatial instances is 10K, and the parameter min _prev is set to 0.6. As shown in Fig.6, as the parameter d increases, the runing time increases.

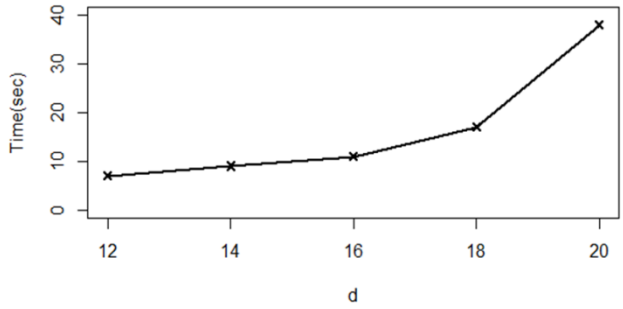

Figure 6. Effct of parameter d

5. CONCLUSION AND FUTURE WORK

In this paper, we discuss the concept of co-location patterns with clustering items and design an algorithm for mining co-location patterns with clustering items. This algorithm is correct. We evaluate the performance of the algorithm by experiments. In order to mining co-location patterns with clustering items more efficiently, we will continue to study it.

ACKNOWLEDGEMENTS

This paper is financially supported by the National Key Research and Development Program of China under Grant numbers 2016YFB0502500, the National Natural Science of China under Grant numbers 41431179, the State Oceanic Administration under Grant numbers 2014#58, GuangXi Natural Science Foundation under Grant numbers 2015GXNSFDA139032, GuangXi Science & Technology Development Program under the Contract number GuiKeHe 14123001-4, GuangXi Key Laboratory of Spatial Information and Geomatics Program under Grant numbers 151400701, 151400712 and 163802512.

REFERENCES

Agarwal, R. and Srikant, R., 1994. Fast algorithms for Mining association rules. In *VLDB*

Huang, Y., Shekhar, S. and Xiong, H., 2004. Discovering Colocation Patterns from Spatial Data Sets: A General Approach. IEEE *Transactions on Knowledge and Data Engineering*, 16(12), 1472-1485

Huang, Y., Pei, J., Xiong, H., 2006. Mining Co-location Patterns with Rare Events from Spatial Data Sets. *GeoInformatica*, 10 (3), 239-260

Huang, Y., Xiong, H., Shekhar S. and Pei, J., 2003. Mining confident co-location rules without a support threshold. *Proceedings of Symposium on Applied Computing*, 497-501

Jiang, Y., Wang, L. and Chen, H., 2010. Discovering both positive and negative co-location rules from spatial data sets. *International Conference on Software Engineering and Data Mining* IEEE, 398-403

Shekhar, S. and Huang, Y., 2001. Co-location Rules Mining:A Summary of Results. *In Proc. Intl. Symposium on Spatio and Temporal Database*

Xiao, X., Xie, X., Luo, Q. and Ma, W., 2008. Density based co-location pattern discovery. *ACM Sigspatial International Symposium on Advances in Geographic Information Systems. DBLP*, 1-10

Yoo, J. S. and Shekhar, S., 2004. A Partial Join Approach for Mining Co-location Patterns. *In Proc. of ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS)*, Washington DC USA, 241-249

Yoo, J. S., Shekhar, S., Celik, M., 2005. A Join-less Approach for Co-location Pattern Mining: A Summary of Results. In the fifth IEEE *International Conference on Data Mining (ICDM ' 05)*, 813-816

Yoo, J. S. and Shekhar, S., 2006. A Join-less Approach for Mining Spatial Co-location Patterns. TKDE, 18(10): 1323-1337
Zhou, G., Song, C. and Schickler, W., 2004. Urban 3D GIS from LIDAR and aerial image data. *Computers and Geosciences*, vol. 30, no. 4, 345-353

Zhou, G. and Wang, L., 2012. Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation. *Transportation Research Part C*, vol. 21, 287–305

Zhou, G. and Wang, L., 2010. GIS and Data Mining to enhance pavement rehabilitation decision-making. *Journal of Transportation Engineering*, vol. 136, no. 4, February, 332-341

Zhou, G., Chen, W. and Kelmelis, J., 2005. A comprehensive study on urban aerial image orthorectification for national mapping program. IEEE *Trans. On Geoscience and Remote Sensing*, vol. 43, no. 9, 2138-2147

Zhou, G., Zhang, R., 2016. Manifold Learning Co-Location Decision Tree for Remotely Sensed Imagery Classification. *Remote Sensing*, 8, 855; doi:10.3390/rs8100855