Geoinformatica (2006) 10: 239–260
DOI 10.1007/s10707-006-9827-8

Mining Co-Location Patterns with Rare Events from Spatial Data Sets

Yan Huang · Jian Pei · Hui Xiong

Abstract A *co-location pattern* is a group of spatial features/events that are frequently co-located in the same region. For example, human cases of West Nile Virus often occur in regions with poor mosquito control and the presence of birds. For co-location pattern mining, previous studies often emphasize the equal participation of every spatial feature. As a result, interesting patterns involving events with substantially different frequency cannot be captured. In this paper, we address the problem of *mining co-location patterns with rare spatial features*. Specifically, we first propose a new measure called the *maximal participation ratio* (maxPR) and show that a co-location pattern with a relatively high maxPR value corresponds to a co-location pattern containing rare spatial events. Furthermore, we identify a weak monotonicity property of the maxPR measure. This property can help to develop an efficient algorithm to mine patterns with high maxPR values. As demonstrated

A preliminary version of the paper appeared as [13].

The research of the second author is supported in part by Natural Sciences and Engineering Research Council of Canada under grant number 312194-05 and National Science Foundation of the United States under grant number IIS-0308001. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agency.

Y. Huang (✉)
Department of Computer Science and Engineering, University of North Texas, Texas, USA
e-mail: huangyan@unt.edu

J. Pei
School of Computing Science, Simon Fraser University, Burnaby, Canada
e-mail: jpei@cs.sfu.ca

H. Xiong
Management Science and Information Systems Department, Rutgers University, Newark, USA
e-mail: hui@rbs.rutgers.edu

by our experiments, our approach is effective in identifying co-location patterns with rare events, and is efficient and scalable for large-scale data sets.

Keywords Spatial data mining · Co-location patterns · Spatial association rules

1 Introduction

Advanced spatial data collecting systems, such as NASA Earth's Observing System (EOS) and Global Positioning System (GPS), have been accumulating increasingly large spatial data sets [8], [12], [18], [24], [25], [30]. For instance, since 1999, more than a terabyte of data has been produced by EOS every day. These spatial data sets with explosive growth rate are considered nuggets of valuable information. The automatic discovery of interesting, potentially useful, and previously unknown patterns from large spatial datasets is being widely investigated via various spatial data mining [16], [23], [24], [29] techniques. Classical spatial pattern mining methods include spatial clustering [22], spatial characterization [9], spatial outlier detection [27], spatial prediction [28], and spatial boundary shape matching [15].

Mining *spatial co-location patterns* [7], [10], [11], [20], [21], [26], [31] is an important spatial data mining task. *A spatial co-location pattern is a set of spatial features that are frequently located together in spatial proximity*. To illustrate the idea of spatial co-location patterns, let us consider a sample spatial data set, as shown in Fig. 1. In the figure, there are various spatial instances with different spatial features that are denoted by different symbols. As can be seen, spatial feature $+$ and $\times$ tend to be located together because their instances are frequently located in spatial proximity.

The problem of mining spatial co-location patterns can be related to various application domains. For example, in location based services, different services are requested by service subscribers from their mobile PDA's equipped with locating devices such as GPS. Some types of services may be requested in proximate geographic area, such as finding the nearest Italian restaurant and the nearest parking place. Location based service providers are very interested in finding what services are requested frequently together and located in spatial proximity. This information can help them improve the effectiveness of their location based recommendation systems where a user requested a service in a location will be recommended a service in a nearby location. Knowing co-location patterns in location based services may also enable the use of pre-fetching to speed up service delivery. In ecology, scientists are interested in finding frequent co-occurrences among spatial features, such as drought, EI Nino, substantial increase/drop in vegetation, and extremely high precipitation.

The previous studies on co-location pattern mining emphasize frequent co-occurrences of *all* the features involved. This marks off some valuable patterns involving rare spatial features. We say *a spatial feature is rare if its instances are substantially less than those of the other features in a co-location*. This definition of "rareness" is relative with respect to other features in a co-location. A feature could be rare in one co-location but not rare in another. For example, if the spatial feature A has 10 instances, the spatial feature B has 20 instances, and the spatial feature C has 10,000 instances. A is not considered a rare feature in the co-location $\{A, B\}$ but it is considered a rare feature in co-location $\{A, C\}$. Of course, a feature with very small number of instances are often rare in many co-location patterns.

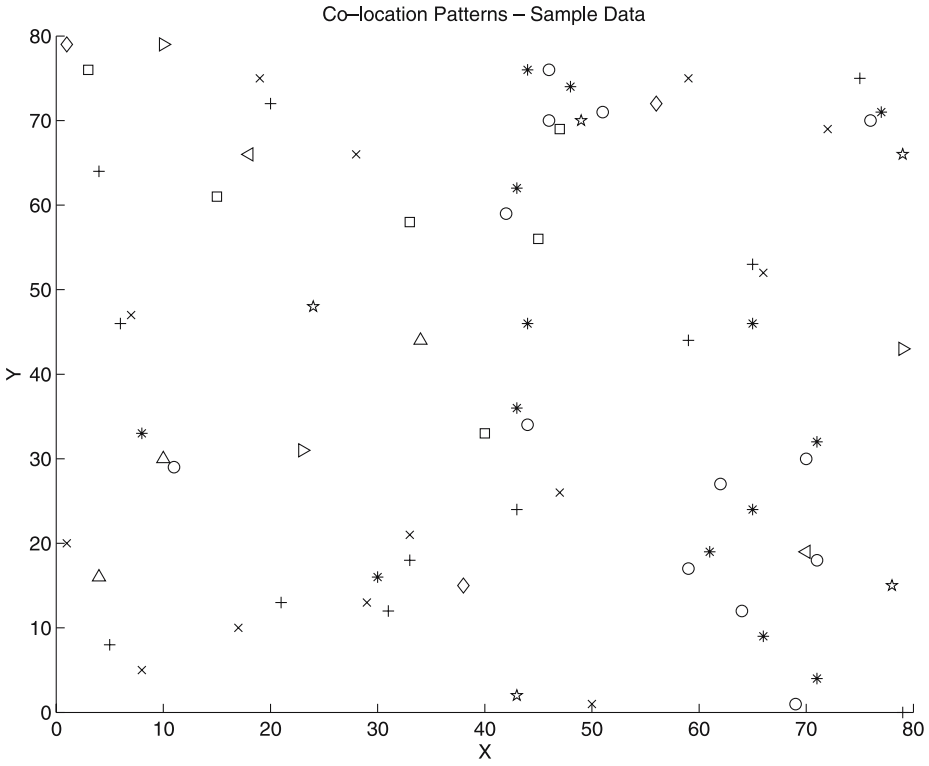


Fig. 1 An illustration of spatial co-location patterns. Shapes represent different spatial feature types. Instances of spatial features in sets {'+', '×'} and {'o', '*'} tend to be located together

In many cases, it is important to capture co-location patterns with rare features. For example, it is believed that human West Nile Virus disease [2] often occurs in regions with poor mosquito control and the presence of birds. The Center for Disease Control has received confirmation from state agencies of 8,219 human cases of West Nile Virus for the year 2003. However, due to numerous locations with poor mosquito control and the presence of birds, we may not find that poor mosquito control and domestic animals are strongly co-located with human West Nile Virus disease using the existing co-location mining methods.

As another example, in a case settled in 1996 [1], *PG&E*'s nearby plant was leaching chromium 6, a rust inhibitor, into the water supply of Hinkley California, and the suit blamed the chemical for dozens of symptoms, ranging from nosebleeds to breast cancer, Hodgkin's disease, miscarriages and spinal deterioration. The prosecutors argued that chromium 6 contaminated water caused nosebleeds, breast cancer, etc. in their nearby region with high probability. Again, this is a typical co-location pattern involving rare spatial features; that is, the spatial event "chromium 6 contaminated water" is rare compared to nose-bleeding.

Therefore, it is necessary to explore new methods to discover co-location patterns with rare spatial features, which is the motivation of this paper. However, the existing co-location mining algorithms [20], [26] have difficulties in identifying such patterns. In general, the challenges of mining spatial co-location patterns with rare spatial features lay in two aspects.

1. *How to identify and measure spatial co-location patterns involving rare spatial features ?*
   Strong interactions involving rare spatial features are often marked off in previous methods, since they require frequent co-occurrences of all features in the co-location patterns. Many measures are based on the measures of *frequency or minimum participation ratio* where rare events are unfavorable.
   Our contributions. In this paper, we propose a novel measure called *maximal participation ratio*, which can incorporate the spatial co-location patterns in the presence of rare spatial features. We show that finding spatial co-locations from spatial data sets with rare spatial features can be achieved by finding co-location patterns with respect to the maximal participation index.
2. *How to mine the patterns involving rare spatial features efficiently?*
   Even though we have a good measure for co-location patterns in the presence of rare spatial features, it is still challenging to find all the patterns efficiently. One dominant obstacle is that *the maximal participation ratio is not monotonic with respect to co-location pattern containment relation*. Thus, the conventional apriori-like pruning technique [4] cannot be applied. Without proper pruning, there could be many possible combinations. Checking them one by one may be computationally prohibitive in many cases.
   Our contributions. In this paper, we study the problem of efficiently mining co-location patterns with rare spatial features systematically. We propose two algorithms. The first algorithm is a rudimentary extension of the apriori-like [4] solution. It uses a very low participation index threshold to prune and use the maximal participation ratio threshold to do a post-processing. It is not efficient since it has to enumerate many patterns.
   Our second algorithm is much more efficient. It exploits an interesting *weak monotonic property* of the maximal participation ratio to push the maximal participation ratio threshold deep into the mining. It achieves good performance in most cases.
   We conduct an extensive performance study to test our methods. The experimental results show that our methods are effective, efficient and scalable for mining large spatial databases.

The remainder of this paper is organized as follows. In Section 2, we review related work. We recall important concepts of association rule mining and compare it with spatial co-location mining in Section 3. Section 4 presents an overview of the co-location pattern mining framework [26]. In Section 5, we introduce the maximal participation ratio. Efficient algorithms for mining co-location patterns with rare features are proposed in Section 6. An extensive performance study is reported in Section 7. Finally, in Section 8, we draw conclusions and suggest future work.

2 Related Work

The previous methods of mining co-location patterns can be divided into two categories, namely the *spatial data mining methods* and the *spatial statistics methods*. They are reviewed briefly in this section.

2.1 Spatial Data Mining Methods

In [26], [31], efficient algorithms were proposed to mine spatial co-location patterns from spatial databases. A set of spatial features form a pattern if, for each spatial feature, at least $s\%$ instances of that feature form a clique with some instance of all the rest features in the pattern for a given neighborhood relationship, such as an Euclidean distance threshold. The parameter $s\%$ is called the participation index. In other words, a set of spatial features form a pattern if whenever a feature of the set is observed, with a probability of at least $s\%$, all other features are also observed in spatial proximity. When the number of objects of different spatial features spans a wide range, the popular features (features with a large number of objects) tend to get a low ratio compared to rare features (features with a small number of instances). In [31], spatial co-location patterns were generalized and expressed by multi-way spatial joins. The space partitioning algorithms were proposed to solve the spatial co-location pattern mining problem. The proposed algorithm is not restricted to a particular interesting measure.

In [20], graphs formed by neighboring spatial instances are partitioned to disjoint parts. A frequency-based pruning technique is developed. This frequency-based pruning method also favors popular spatial features. A clustering-based map overlay approach [10], [11] treats every spatial attribute as a map layer and considers spatial clusters (regions) of point-data in each layer as candidates for mining associations. Given X and Y as sets of layers, a clustered spatial association rule is defined as $X \Rightarrow Y(CS\%, CC\%)$, for $X \bigcap Y = \emptyset$, where $CS\%$ is the clustered support, defined as the ratio of the area of the cluster (region) that satisfies both X and Y to the total area of the region S under investigation, and $CC\%$ is the clustered confidence, which can be interpreted as $CC\%$ of the areas of clusters (regions) of X intersect with areas of clusters(regions) of Y. However, instances of rare spatial features, e.g., chromium 6 populated water sources, do not always form clusters or regions.

The reference feature centric model proposed in [17] enumerates proximity neighborhoods to "materialize" a set of transactions around instances of a user specified reference spatial feature. Transactions are created around instances of one user-specified spatial feature. The association rules are derived using the apriori algorithm [4]. The rules found are all related to the reference feature. The support based apriori pruning marks off co-location patterns with rare spatial features.

Munro et al. [21] described the need for mining complex relationships in spatial data including multi-feature colocation, self-colocation, one-to-many relationships, self-exclusion and multi-feature exclusion.

2.2 Spatial Statistics Methods

In spatial statistics, some dedicated techniques such as cross k-functions with Monte Carlo simulations [7], mean nearest-neighbor distance, and spatial regression models

[6] have been developed to test the co-location of two spatial features and find pairs of co-located spatial features. However, the Monte Carlo simulation could be expensive. Another approach is to arbitrarily partition the space into a lattice. For each cell of the lattice, count the number of instances of each spatial feature. Pairwise correlation of spatial features could be found by tests such as χ^2 [7] or using classic association rule mining algorithms such as apriori [4] by treating each cell as a transaction. Arbitrary partitioning may loss neighboring instances across borders of cells. Both the cross k-function and the pair wise correlation cannot be easily extended to the cases with more than two spatial features.

To the best of our knowledge, this is the first systematic study on mining co-location patterns with rare features in the spatial context.

3 Association Rule Mining

Since spatial co-location pattern mining resembles association pattern mining [3] in many aspects, we review the basic concepts of association rules in this section.

Since its introduction [3], the problem of mining association rules from large databases has been the subject of numerous studies. The *association rule mining problem* is defined as follows.

Let $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$ be a set of m *items*. Let $TD = \{T_1, T_2, \ldots, T_n\}$ be a *transactional database* where $T_i (i \in [1, n])$ is a *transaction* which is a subset of items in $\mathcal{I}$. For an *itemset* $Y \subseteq \mathcal{I}$, the *support* of Y is the number of transactions containing Y in TD, i.e., $sup(Y) = |\{T_i | T_i \supseteq Y\}|$. Y is of size k if $|Y| = k$.

The *confidence* of an *association rule* in the form of $X \rightarrow Y$, where $X \cap Y = \emptyset$, is the ratio of the support of $X \cup Y$ versus the support of X. Itemset Y is a *frequent pattern* if the support of Y is no less than a *minimum support threshold* specified by user. We compare and contrast frequent pattern mining and spatial co-location mining in Table 1.

The support of itemsets has a downward closure property (sometimes called the apriori property): *the support of Y is no less than the support of any superset of Y*. Because of the downward closure property of the support, a generate-and-test mining paradigm was employed by the apriori algorithm proposed in [4]. This approach generates candidates of size $(k + 1)$ items set based on the size k frequent itemsets. The set of size $(k + 1)$ candidates includes all and only those itemsets of size $(k + 1)$ whose size k subsets are all frequent. False candidates are pruned by scanning the transactions before the next iteration.

Table 1 Comparison of frequent pattern mining and spatial co-location mining

Frequent pattern mining	Co-Location mining
Item	Spatial feature
Item set	Spatial feature set
Frequent pattern	Co-location pattern
Support	Spatial interestingness measures
Transactional database	Spatial database

4 Co-Location Patterns in Spatial Databases

In this section, we review a framework of mining co-location patterns, since our proposed solution in this paper is based on this model. The framework was proposed in [26] and is based on the *participation index*. We will point out why such a framework still may miss some co-location patterns involving rare spatial features. In the next section, we will extend the framework to mine co-location patterns with rare spatial features.

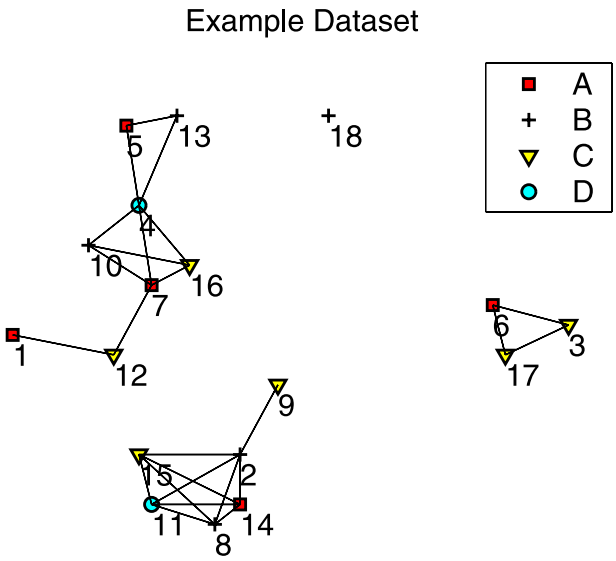
For a spatial data set S, let $F = \{f_1, \ldots, f_k\}$ be a set of *boolean spatial features*. Let $\boldsymbol{i} = \{i_1, \ldots, i_n\}$ be a set of n instances in S, where each instance is a vector ⟨instance-id, location, spatial features⟩. The spatial feature f of instance i is denoted by $i.f$. We assume that the spatial features of an instance are from F and the location is within the spatial framework of the spatial database. Furthermore, we assume that there exists a neighborhood relation $\mathcal{R}$ over pairwise instances in S.

Example 1: (A Spatial Data Set) Figure 2 shows a spatial data set with a spatial feature set $F = \{A, B, C, D\}$, which will be used as the running example in this paper. Objects with various shape represent different spatial features, as shown in the legend. Each instance is uniquely identified by its instance-id. We have 18 instances in the database. ■

The objective of co-location pattern mining is to find frequently co-located subsets of spatial features. For example, a co-location *{traffic jam, police, car accident}* means that a traffic jam, police, and a car accident frequently occur in a nearby region.

To capture the concept of "nearby," the concept of user-specified neighbor-sets was introduced. A *neighbor-set* L is a set of instances such that all pairwise locations in L are neighbors. A *co-location pattern* (or just pattern for short) C is a set of spatial features, i.e., $C \subseteq F$. A neighbor-set L is said to be a *row instance* of co-location

Fig. 2 An example dataset



pattern C if every feature in C appears as a feature of an instance in L, and there exists no proper subset of L does so. We denote all row instances of a co-location pattern C by $rowset(C)$.

Example 2: (Neighbor-set, row instance and rowset) In Fig. 2, the neighborhood relation $\mathcal{R}$ is defined based on Euclidean distance. Two instances are neighbors if their Euclidean distance is less than a user specified threshold. Neighboring instances are connected by edges. For instance, {3, 6, 17}, {4, 5, 13}, and {4, 7, 10, 16} are all neighbor-sets because each set forms a clique. Here, we use the instance-id to refer to an object in Fig. 2. Additional neighbor-sets include {6, 17}, {3, 6}, {2, 15, 11, 14}, and {2, 15, 8, 11, 14}.

$\{A, B, C, D\}$ is a co-location pattern. The neighborhood-set {14, 2, 15, 11} is a row instance of the pattern $\{A, B, C, D\}$ but the neighborhood-set {14, 2, 8, 15, 11} is not a row instance of co-location $\{A, B, C, D\}$ because it has a proper subset {14, 2, 15, 11} which contains all the features in $\{A, B, C, D\}$.

Finally, the $rowset(\{A, B, C, D\}) = \{\{7, 10, 16, 4\}, \{14, 2, 15, 11\}, \{14, 8, 15, 11\}\}$. ■

For a co-location rule $R: A \rightarrow B$, the *conditional probability* $cp(R)$ of R is defined as

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

In words, the conditional probability is the probability that a neighbor-set in $rowset(A)$ is part of a neighbor-set in $rowset(A \cup B)$. Intuitively, the conditional probability p indicates that, whenever we observe the occurrences of spatial features in A, the probability to find occurrence of B in a nearby region is p.

Example 3: (Conditional probability) In Fig. 2, based on the Euclidean distance relation $\mathcal{R}$ as described in Example 2,

$$rowset(\{A, B, C, D\}) = \{\{7, 10, 16, 4\}, \{14, 2, 15, 11\}, \{14, 8, 15, 11\}\},$$

and

$$rowset(\{A, B\}) = \{\{7, 10\}, \{14, 2\}, \{5, 13\}, \{14, 8\}\}.$$

Since $|rowset(\{A, B\})| = 4$, only 3 rows of $\{A, B\}$ satisfy the subset condition, i.e., row {7, 10} of $\{A, B\}$ is a subset of row {7, 10, 16, 4} of $\{A, B, C, D\}$, row {14, 2} of $\{A, B\}$ is a subset of row {14, 2, 15, 11} of $\{A, B, C, D\}$ and row {14, 8} of $\{A, B\}$ is a subset of row {14, 8, 15, 11} of $\{A, B, C, D\}$, the conditional probability $cp(\{A, B\} \rightarrow \{C, D\}) = \frac{3}{4} = 75\%$. ■

Given a spatial database S, to measure how a spatial feature f is co-located with other features in co-location pattern C, a *participation ratio* $pr(C, f)$ can be defined as

$$pr(C, f) = \frac{|\{r | (r \in S) \wedge (r.f = f) \wedge (r \text{ is in a row instance of } C)\}|}{\{r | (r \in S) \wedge (r.f = f)\}|}$$

In words, a feature f has a partition ratio $pr(C, f)$ in pattern C means wherever the feature f is observed, with probability $pr(C, f)$, all other features in C are also observed in a neighbor-set.

In [26], a *participation index* was proposed to measure how all the spatial features in a co-location pattern are co-located. For a co-location pattern C, the participation index $PI(C) = \min_{f \in C}\{pr(C, f)\}$. In words, wherever any feature in C is observed, with a probability of at least $PI(C)$, all other features in C can be observed in a neighbor-set. A high participation index value indicates that the spatial features in a co-location pattern likely occur together. The participation index was proposed because in spatial application domain there are no natural "transactions" and thus "support" is not well-defined.

Given a user-specified *participation index threshold min_prev*, a co-location pattern is called *prevalent* if $PI(C) \geq min\_prev$.

Example 4: (Participation ratio and participation index) To find the participation index $PI(\{A, B, C, D\})$ of pattern $\{A, B, C, D\}$, we first identify the rowsets of $\{A, B, C, D\}$ as shown in Example 2, i.e., $\{\{7, 10, 16, 4\}, \{14, 2, 15, 11\}, \{14, 8, 15, 11\}\}$.

Among all the five instances of A, two of them, namely 7 and 14, have B, C and D in a neighbor-set. So the participation ratio $pr(\{A, B, C, D\}, A) = \frac{2}{5}$. Similarly, we can have $pr(\{A, B, C, D\}, B) = \frac{3}{5}$, $pr(\{A, B, C, D\}, C) = \frac{2}{6} = \frac{1}{3}$, and $pr(\{A, B, C, D\}, D) = \frac{2}{2} = 1$. Taking the minimal of all the ratios, the participation index $PI(\{A, B, C, D\})$ of co-location $\{A, B, C, D\}$ is $\frac{1}{3}$. ■

As shown below, both the participation ratio and the participation index are monotonic with respect to the size of co-location patterns.

Lemma 1: (Monotonicity of participation ratio and participation index [26]) Let C and C' be two co-location patterns such that $C \subset C'$. Then, for each feature $f \in C$, $pr(C, f) \geq pr(C', f)$. Furthermore, $PI(C) \geq PI(C')$.

Proof: To have the first claim in the lemma, we only need to show that for a spatial feature $f \in C$,

$$|\{r|(r \in S) \wedge (r.f = f) \wedge (r \text{ is in a row instance of } C)\}| \geq$$
$$|\{r|(r \in S) \wedge (r.f = f) \wedge (r \text{ is in a row instance of } C')\}|$$

Since $C \subset C'$, every row instance of C' contains a subset of instances which is a row instance of C. Thus, the inequality holds.

The second claim follows the fact that $PI(C) = \min_{f \in C}\{pr(C, f)\} \geq \min_{f \in C}\{pr(C', f)\} \geq \min_{f \in C'}\{pr(C', f)\} = PI(C')$. ■

Based on Lemma 1, a level-by-level, iterative apriori-like algorithm was developed in [26] to find the complete set of prevalent patterns from a spatial database. For details of the algorithm, please refer to [26].

It is interesting to note that, in the above prevalent co-location pattern mining framework, some co-location patterns involving rare spatial features may be unfortunately missed.

Example 5: (A Co-location Pattern for West Nile Disease) Let us consider the co-location pattern $C = \{West\ Nile,\ poor\ mosquito\ control,\ domestic\ animal\}$. Suppose participation ratios

$$pr(C, \text{West Nile }) = 85\%,$$

$$pr(C, \text{poor mosquito control}) = 10\%$$

and

$$pr(C, \text{ domestic animal}) = 1\%.$$

Then, $PI(C) = \min\{85\%, 10\%, 1\%\} = 1\%$. As can be seen, even though West Nile is strongly co-located with poor mosquito control and domestic animal, unfortunately, the whole co-location pattern is weak in the term of participation index because the West Nile is rare compared to poor mosquito control and domestic animals. ■

Can we extend the framework to mine such patterns even though their participation index values are low? In other words, *can we mine co-location patterns with rare spatial features?* We will address this issue in the next two sections.

5 Maximal Participation Ratio

There is one important observation about co-location patterns with rare spatial features, "*even though the participation index of the whole pattern could be low, there must be some spatial feature(s) with high participation ratio(s)*." In Example 5, in pattern $P = \{West\ Nile,\ poor\ mosquito\ control,\ domestic\ animal\}$, the participation index is low, since West Nile disease are rare compared to poor mosquito control and domestic animals. However, the participation ratio of "*West Nile Virus*" in the pattern is high.

The above observation motivates our extension of the participation index framework. For a co-location pattern C, we define the *maximal participation ratio* as $maxPR = \max_{f \in C}\{pr(C, f)\}$. In words, a high maximal participation ratio value indicates that there are some spatial features strongly imply the pattern.

In general, given a co-location pattern $C = \{f_1, \ldots, f_k\}$, we sort all spatial features in C in the participation ratio descending order. Without loss of generality, for a given *minimum maximal participation ratio threshold min_maxPR*, suppose for $i \in [1, l]$ $pr(C, f_i) \geq min\_maxPR$, where $1 \leq i \leq l \leq k$ and l is the last spatial feature that has participation ratio above the user given threshold. The output of the co-location mining with rare spatial features will be in the form of $\langle C = \{f_1, \ldots, f_k\}, l\rangle$. Then, we can say that *if a spatial feature* f_i $(1 \leq i \leq l)$ *is observed in some location, then the probability of observing all other spatial features in* $C - \{f_i\}$ *in a neighbor-set is at least* $maxPR(C)$.

Given a *minimum maxPR threshold min_maxPR*, the problem of *mining co-location patterns with rare spatial features in a spatial database is to find the complete set of co-location patterns C such that* $maxPR(C) \geq min\_maxPR$.

In general, every pattern that is significant in participation index is also significant in maximal participation ratio. In other words, mining co-location patterns with rare

Table 2 Rowsets, *PI*s and *maxPR*s of co-locations of dataset in Fig. 2

ID	Co-loc	Rowset	pr	PI	max PI
1	{A}	{{1},{5},{6},{7},{14}}	{1}	1	1
2	{B}	{{2},{8},{10}, {13},{18}}	{1}	1	1
3	{C}	{{3},{9},{12}, {15},{16},{17}}	{1}	1	1
4	{D}	{{4},{11}}	{1}	1	1
5	{A,B}	{{5,13},{7,10},{14,2},{14,8}}	{4/5,4/5}	4/5	4/5
6	{A,C}	{{1,12},{6,3},{6,17},{14,15},{7,16}}	{4/5,5/6}	4/5	5/6
7	{A,D}	{{5,4},{14,1},{7,4}}	{3/5,2/2}	3/5	1
8	{B,C}	{{2,9},{2,15},{8,15},{10,16}}	{3/5,3/6}	1/2	3/5
9	{B,D}	{{2,11},{8,11},{10,4},{13,4}}	{4/5,2/2}	4/5	1
10	{C,D}	{{15,11},{16,4}}	{2/6,2/2}	1/3	1
11	{A,B,C}	{{7,10,16},{14,2,15},{14,8,15}}	{2/5,3/5,2/6}	1/3	3/5
12	{A,B,D}	{{5,13,4},{7,10,4},{14,2,11},{14,8,11}}	{3/5,4/5,2/2}	3/5	1
13	{A,C,D}	{{7,16,4},{14,15,11}}	{2/5,2/6,2/2}	2/5	1
14	{B,C,D}	{{2,15,11},{10,16,4},{8,15,11}}	{3/5,2/6,2/2}	1/3	1
15	{A,B,C,D}	{{7,10,16,4},{14,2,15,11},{14,8,15,11}}	{2/5,3/5,2/6,2/2}	1/3	1

spatial features using the maximal participation ratio measure will find all prevalent patterns as a subset.

While the extension of participation index to maximal participation index is intuitive, there is no easy way to extend the existing level-by-level apriori-like algorithm [4] to mine patterns with respect to a maximal participation ratio threshold. The dominant obstacle is that *maximal participation ratio is not monotonic with respect to the pattern containment relation*, as shown in the following example.

Example 6: (Maximal participation ratio is not monotonic) In Fig. 2, the set of spatial features $\{B, C\} \subset \{A, B, C\}$. However, $maxPR(\{B, C\}) = \max\{\frac{3}{5}, \frac{3}{6}\} = 60\% \leq maxPR(\{B, C, D\}) = \max\{\frac{3}{5}, \frac{2}{6}, \frac{2}{2}\} = 100\%$! (Please see Table 2 for the rowsets and the *maxPR*'s). ■

Now, the challenge becomes *how we can push the maximal participation ratio threshold to prune the search space*. That is the topic of the next section.

6 Algorithms

In this section, we will develop efficient algorithms for mining co-location patterns from spatial databases with rare spatial features using maximal participation ratio measure. We propose two methods. The first method is a rudimentary extension of the Apriori algorithm [4]. The second method is based on an interesting weak monotonic property of the maximal participation index.

6.1 A Rudimentary Algorithm

As shown in the previous section, the maximal participation ratio is not monotonic. Thus, we cannot apply the apriori-like pruning directly. In many applications, very rare events could be just noise. Thus, we may in fact have a minimum prevalent threshold *min_prev* and a minimum maximal participation ratio threshold *min_maxPR* such that we only want to find patterns P with $PI(P) \geq min\_prev$ and $maxPR(P) \geq min\_maxPR$.

Based on this observation, we can develop an apriori-like algorithm as follows. We use the minimum prevalent threshold *min_prev* to do apriori-like pruning, then filter out patterns failed the maximal participation ratio threshold by a post-processing.

To ease the presentation, we call a co-location pattern with k spatial features a *k-pattern*. We assume that the spatial features in a k-pattern C is ordered and indexed by their positions in the co-location pattern, i.e., f_i means the ith spatial feature in C. The algorithm, called *Min–Max*, is presented in Fig. 3.

The geometric algorithm is used in step 1 to generate length-2 candidates, since all singleton co-location patterns have both participation index and max participation index equal to 1, and do not need to be checked. Spatial join methods utilizing minimal rectangle bounding box, such as the well known plane sweep [5], space partition [14], and tree matching [19], can be used.

Example 7: (Algorithm Min–Max) Suppose $min\_prev = 0$ and $min\_maxPR = 0.85$, let us show one iteration of the algorithm from 2-patterns to 3-patterns for the dataset in Fig. 2.

Input: A spatial database S, a neighborhood relation $\mathcal{R}$, a minimum prevalent threshold *min_prev*, and a minimum maximal participation index threshold *min_maxPR*.

Output: Co-location patterns P such that $PI(P) \geq min\_prev$ and $maxPR(P) \geq min\_maxPR$.

Method:

1. let $k = 2$; generate C_2, the set of candidate 2-patterns and their rowsets, by geometric methods;
2. for each $C \in C_k$ calculate $PI(C)$ and $maxPR(C)$ from C's rowset $rowset(C)$;
3. let P'_k be the subset of C_k such that for each $P \in P'_k$, $PI(P) \geq min\_prev$;
4. let P_k be the subset of P'_k such that for each $P \in P_k$, $maxPR(P) \geq min\_maxPR$;
5. generate the set C_{k+1} of candidate $(k+1)$-patterns, a co-location pattern P with $(k+1)$ spatial features is in C_{k+1} if and only if for each feature $f \in P$, $(P - \{f\}) \in P'_k$;
6. if $C_{k+1} \neq \emptyset$, let $k = k + 1$, go to Step 2;
7. output $\cup_i P_i$ ■

Fig. 3 Algorithm Min–Max

From Table 2 , we have

$$P'_2 = \{\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}, \{C, D\}\}$$

and

$$P_2 = \{\{A, C\}, \{A, D\}, \{B, D\}, \{C, D\}\}.$$

From those rowsets, it is straightforward to calculate their *PI*s and *maxPRs*. The algorithm generates candidate 3-patterns $C_3 = \{\{A, B, C\}, \{A, B, D\}, \{A, C, D\}, \{B, C, D\}\}$ from P'_2. Then the rowsets of the candidates are generated by joining the rowsets of the two 2-patterns. For example, the rowset of {A,B} joins the rowset of {A,C} to produce the rowset of $\{A, B, C\}$. At the end of this iteration, we have the set of candidate 3-patterns C_3 and their rowsets. C_3 is not empty. So, we start the next round from step 2 in the algorithm. ■

When the minimum prevalent threshold is set to 0, the algorithm can find the complete set of patterns. If *min_prev* is over 0, some patterns with high maximal participation ratio but low prevalence may be missed. In Example 7, if *min_prev* is set to 0.45, $PI(\{A, D\}) = 0.4$ and $\{A, D\}$ is not in P'_2. $\{A, C\}$ and $\{A, D\}$ will not join to produce candidate $\{A, C, D\}$, though $\max PI(\{A, C, D\}) = 1 \geq min\_max PR$.

One advantage of the Min–Max algorithm is that the user can specify the prevalence of patterns she wants to see by the *min_prev* value. The major disadvantage of the algorithm is that, if a user wants to find the complete answer, the algorithm has to generate a huge number of candidates and test them, even though the maximal participation ratio threshold $min\_max PR$ is high.

6.2 Pruning by a Weak Monotonic Property

Is there any property of the maximal participation ratio we can use to get efficient algorithms for co-location pattern mining with rare features?

Let us re-examine Example 5. Pattern $P =$ {West Nile, poor mosquito control, domestic animal} has three proper subsets such that each subset has exactly 2 features. Feature *West Nile* has a high participation ratio, and it participates in two out of the three subsets. Since the participation ratio is monotonic (Lemma 1), the maximal participation ratio values of the two proper subsets containing *West Nile* must be higher or equal to that of P. In other words, at most one 2-subpattern of P can have a lower maximal participation ratio value.

The above observation can be generalized to a pattern with l features. Thus, we have the following *weak monotonic* property.

Lemma 2: (Weak monotonicity) Let P be a k-co-location pattern. Then, there exists at most one $(k-1)$-subpattern P' such that $P' \subset P$ and $max PR(P') < max PR(P)$.

Proof: Let $f_j \in P$ be a spatial feature whose participation ratio is maximal in P. For all $(k-1)$-pattern P' such that $(P' \subset P) \wedge (P' \neq P/\{f_j\})$, P' contains f_i and $f_i \in P' \cap P$. Based on Lemma 1, $max PR(P') \geq pr(P', f_j) \geq pr(P, f_j) = max PR(P)$. In other words, only one $(k-1)$-subpattern of P, i.e., $P/\{f_j\}$, is possible to have a lower maximal participation index value than P does. ■

Based on the above weak monotonic property, if a k-pattern is above the maximal participation ratio threshold, then at least $(k-1)$ out of its k subpatterns with $(k-1)$ features are above the maximal participation ratio threshold. Therefore, we can revise the candidate generation process, such that only a k-pattern having at most one $(k-1)$-subpattern below the minimum maximal participation ratio min_maxPR threshold should be generated. The idea is illustrated in the following example.

Example 8: (Candidate generation using weak monotonicity) Suppose the maximal participation ratio values of $\{A, B, C\}$, $\{A, C, D\}$ and $\{B, C, D\}$ are all over the threshold min_maxPR, but that of $\{A, B, D\}$ is not. We still should generate a candidate $P = \{A, B, C, D\}$, since it is possible that $maxPR(P)$ passes the threshold.

To achieve this, we need a systematic way to generate the candidates. Please note that, in apriori, for the above example, $\{A, B, C, D\}$ is generated only if $\{A, B, C\}$ and $\{A, B, D\}$ (differ only in their last spatial feature) are both frequent. However, in the co-location pattern mining with rare spatial features using maximal participation ratio measure, it is possible that $\{A, B, D\}$ is below the given threshold min_maxPR while $\{A, B, C, D\}$ is above the threshold min_maxPR.

In general, for two co-location patterns P and P' from the set P_k of k-patterns above threshold min_maxPR, i.e., $P \in P_k$ and $P' \in P_k$, P and P' can be joined to generate a candidate $(k+1)$-pattern in C_{k+1} if and only if P and P' have one different feature in the last two features. For example, even $\{A, B, D\}$ is below threshold min_maxPR, candidate $\{A, B, C, D\}$ can be generated by $\{A, B, C\}$ and $\{A, C, D\}$ since they have the common feature C in their last two features, i.e., they differ one spatial feature in their last two spatial features. ■

We will illustrate the correctness of the above candidate generation method in Lemma 3 and Example 9. Also, with the revised candidate generator, the mining algorithm is presented in Fig. 4.

The algorithm does not need a minimum prevalence threshold but still finds all co-location patterns with maximal participation index above threshold min_maxPR.

To make sure the candidate generation does not miss any co-location, we need to prove that the candidate $(k+1)$-patterns C_{k+1} generated by the maxPrune algorithm

Input: A spatial database S, a neighborhood relation $\mathcal{R}$, a minimum maximal participation ratio min_maxPR.
Output: Co-location patterns P such that $maxPR(P) \geq min_$maxPR.
Method:

1. let $k = 2$; generate C_2, the set of candidate 2-patterns and their rowsets, by geometric methods;
2. For each $C \in C_k$ calculate $maxPR(C)$ from C's rowset $rowset(C)$; Let P_k be the subset of C_k such that for each $P \in P_k$, $maxPR(P) \geq min\_maxPR$;
3. generate C_{k+1}, the set of candidates $(k+1)$-patterns, as illustrated in Example 8 ; if $C_{k+1} \neq \emptyset$, let $k = k+1$, go to Step 2;
4. output $\cup_i P_i$ ■

Fig. 4 Algorithm maxPrune

is a superset of the actual $(k+1)$-patterns P_{k+1}. This is proved in the following lemma.

Lemma 3: Let P be a k-pattern above given threshold min_maxPR ($k \geq 3$). Then, there exist two $(k-1)$ patterns P_1 and P_2 such that (1) $P_1 \subset P$, $P_2 \subset P$, (2) P_1 and P_2 share their first $k-2$ features, (3) P_1 and P_2 share either the kth or the $(k-1)th$ feature or the $(k-2)th$ feature in P but not any two of them, and (4) both P_1 and P_2 are above threshold min_maxPR.

Proof: The three length-$(k-1)$ sub-patterns: $P/\{f_{k-2}\}$,$P/\{f_{k-1}\}$, and $P/\{f_k\}$ share their first $k-2$ features and pairwise share one feature (kth feature, $(k-1)$th feature, or $(k-2)th$ feature of P) in their last two features but not two of the features in $\{f_{k-2}, f_{k-1}, f_k\}$. Following Lemma 2, P has at most one length-$(k-1)$ subpattern P' which is below threshold min_maxPR. Thus, if any one of the three below threshold min_maxPR, we still have the other two length-$(k-1)$ patterns as stated in the Lemma. ■

Lemma 3 guarantees that if we generate size k candidate patterns by joining any two $(k-1)$ patterns which differ in one feature in their last two features, we will not miss any co-location patterns above threshold min_maxIP.

Example 9: (Algorithm maxPrune) Suppose $min\_maxPR = 0.85$. Initially, all singleton co-location patterns are qualified since they have $maxPR = 1$. A general geometric method is used to generate candidate two-patterns and their rowsets. From their rowsets, we calculate their $maxPR$. Only the co-location patterns in

$$P_2 = \{\{A, C\}, \{A, D\}, \{B, C\}, \{C, D\}\}$$

are above min_maxPR threshold.

Then, we generate candidates 3-patterns. In detail, $\{A, C\}$ joins $\{A, D\}$, $\{A, C\}$ joins $\{B, C\}$, $\{A, C\}$ joins $\{C, D\}$, $\{A, D\}$ joins $\{C, D\}$, and $\{B, C\}$ joins $\{C, D\}$ to generate candidate 3-patterns. After duplicate elimination, we have

$$C_3 = \{\{A, B, C\}, \{A, B, D\}, \{A, C, D\}\{B, C, D\}\}.$$

The rowsets of the candidates are generated by joining the rowsets of the two two-patterns leading to the candidate.

We go back to step 2. From their rowsets, we calculate the maximal participation ratio values for the candidate 3-patterns. We get $P_3 = \{\{A, B, D\}, \{A, C, D\}$ $\{B, C, D\}\}$. The patterns in P_3 are above threshold min_maxPR. Then, we generate the candidate 4-patterns. In detail, $\{A, B, D\}$ joins $\{A, C, D\}$ because they differ by one feature in their last two features. We thus generate candidate four-pattern $C_4 = \{A, B, C, D\}$, as illustrated in Example 8. Rowsets of $\{A, C, D\}$ and $\{B, C, D\}$ are joined to produce the rowset of $\{A, B, C, D\}$. Its $maxPR$ is calculated and it is above the threshold min_maxPR.

The algorithm proceeds similarly. It can be verified that $C_5 = \emptyset$ and thus the algorithm stops. ■

Compared to the min–max algorithm, the maxPrune algorithm does not need any minimum prevalence threshold and finds the complete set of co-locations above the

minimum maximal participation index threshold min_maxPR with any prevalence. In the process of mining the complete set of these co-locations, the maxPrune algorithm generates much less candidate co-location patterns compared to that of min–max with $min\_prev = 0$, and thus lowers down the costs of expensive rowset generation and test dramatically.

7 Experimental Results

In this section, we present extensive experiments to evaluate the performance of two algorithms: Min–Max and maxPrune. Specifically, we test three cases: (1) data sets which contains no co-location patterns involving rare spatial features, (2) data sets containing patterns involving rare spatial features, and (3) large data sets.

7.1 The Experimental Setup

Our experiments were performed on synthetic data sets. We developed a data generator for generating synthetic data. Our data generator is similar to the one used in [4], with some extensions to produce spatial data sets. The major parameters for generating synthetic data sets are illustrated as follows. For a synthetic data set $I100k.C10.R50$, we generate 100 k instances (denoted as $I100k$). There are up to 50 co-location patterns containing rare features with high participation ratio but very low prevalence (denoted as $R50$). We achieve this by binding a spatial feature to a pre-generated potential co-location pattern, and making those bound spatial features not prevalent. The number of features in a co-location pattern yields to a Poisson distribution, while the mean is 10 (denoted as $C10$). For all generated data sets, the total number of features is 100 and the total number of pre-generated potential co-location patterns is 500. A summary of the parameter settings used to generate the synthetic data sets is presented in Table 3.

We implemented algorithms min–max and maxprune using C++ and all experiments were performed on a Pentium III 550 MHz PC machine with $4G$ MB main memory, running Linux Redhat 6.1 operating system. Due to the space limit, we only report the results on some representative data sets.

Table 3 Parameter settings

Name	$\|I\|$	$\|C\|$	$\|R\|$	Size MB
$I100$ K.C5.R0	100K	5	0	1.86
$I100$ K.C5.R50		5		
$I100$ K.C10.R50	100K	10	50	1.86
$I100$ K.C15.R50		15		
$I250$ K.C5.R50		5		
$I250$ K.C10.R50	250K	10	50	4.8
$I250$ K.C15.R50		15		
$I750$ K.C5.R50		5		
$I750$ K.C10.R50	750K	10	50	14.6
$I750$ K.C15.R50		15		
$I1$ M.C5.R50		5		
$I1$ M.C10.R500	1M	10	50	19.6
$I1$ M.C15.R50		15		

7.2 The Performance Comparison on Data Sets in Which no Co-Location Patterns have High Maxpi Values

We first evaluate the performance of Min–Max and maxPrune algorithms on mining prevalent co-location patterns from spatial data sets in which there are no co-location patterns with high maxPR values. This experiment was conducted on the data set $I100kC5R0$. The parameter $R0$ in this data set indicates that no co-location patterns have relatively high maxPR values. In other words, the *PI* and *maxPR* do not vary too much. Referred to the discussion in the previous sections, this setting favors the Min–Max algorithm.

In the experiment, we varied maxPR thresholds from 2.5 to 4%. Figure 5(a) shows the runtime of maxPrune and Min–Max at different *maxPR* thresholds. For the Min–Max algorithm, we chose the *min_prev* values as 0, 0.05, 0.5 and 2.5%, respectively. Only when the *min_prev* value was set to 0%, Min–Max finds all the co-location patterns with maxPR values above the maxPR threshold. Because the Min–Max algorithm has to generate all co-location patterns with participation index values above *min_prev*, the run time of Min–Max is not sensitive to the change of maxPR thresholds. In contrast, the runtime of maxPrune increases as the maxPR threshold decreases.

When $min\_prev > 0$, the Min–Max algorithm can potentially miss some interesting co-location patterns. When two algorithms generate the same set of co-location patterns, i.e., if *min_prev* is equal to zero for the Min–Max algorithm, Min–Max outperforms maxPrune only when the maxPR threshold is lower than 3.2%. In such an extreme region, most co-location patterns have high maxPR values, and algorithm maxPrune has a heavier overload on candidate generation than Min–Max has. For all other parameter regions, maxPrune outperforms Min–Max.

Finally, Fig. 5(b) shows the number of co-location patterns with high maxPR values identified at different maxPR thresholds. As can be seen, only if a very small *min_prev* threshold is specified, the Min–Max algorithm can generate comparable results as maxPrune does. However, the computation performance of the Min–Max algorithm degrades a lot when the *min_prev* threshold is small.

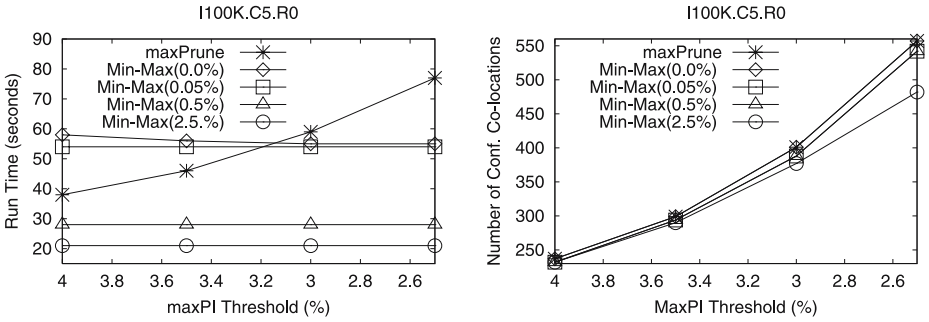


Fig. 5 (a) The runtime of Min–Max and Maxprune on data sets in which there are no co-location patterns with high maxPI values. (b) The number of Co-locations identified on data sets in which no Co-location has high maxPI values

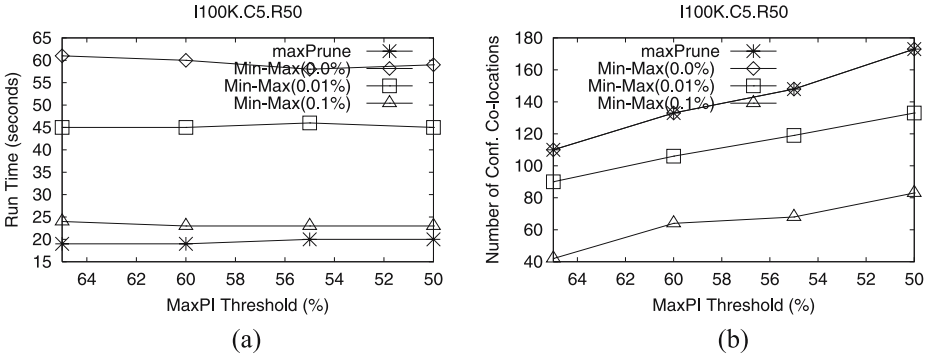


Fig. 6 (a) The runtime of Min–Max and maxPrune on data sets containing co-locations with high maxPR values. (b) The number of co-location patterns identified on data sets containing co-locations with high maxPI values

7.3 Performance Comparison on Data Sets Containing Co-Location Patterns with High Maxpi Values

Here, we compare the performance of maxPrune and Min–Max on a data set (*I*100K.C5.R50) in which there are many co-location patterns with high maxPR values but low prevalence. In this experiment, we identified many co-location patterns with relatively high maxPR values, say above 50%.

Figure 6(a) shows the runtime of Min–Max and maxPrune on data set *I*100K.C5.R50. As can be seen, the runtime of Min–Max dramatically increases with the decrease of participation index thresholds. In contrast, the runtime of maxPrune is not affected by the change of participation index thresholds and is much smaller than that of Min–Max with respect to different MaxPI thresholds. In addition, Fig. 6(b) shows that the number of co-location patterns identified by Min–Max decreases with the increase of participation index thresholds. In other words, for the min–max algorithm, there is a trade-off between the efficiency and the completeness of results. However, the maxPrune algorithm does not have such a dilemma situation.

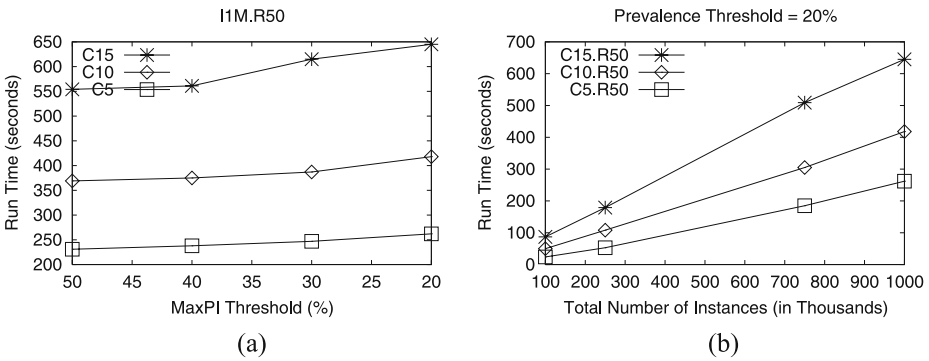


Fig. 7 (a) Scalability of the maxPrune algorithm w.r.t. maxPR. threshold (b) Scalability of the maxPrune algorithm w.r.t. number of instances

7.4 The Scalability of Maxprune

In this subsection, we first evaluate the scalability of the maxPrune algorithm with respect to maxPR thresholds. Figure 7(a) shows the runtime of the maxPrune algorithm on data set I1M.C15.R50. As can be seen, the runtime of the maxPrune algorithm almost linearly increases with the decrease of maxPR thresholds. Another observation is that more features on average are in a pattern, the long the runtime we will have. Also, Fig. 7(b) shows the scalability of maxPrune in terms of the number of instances in spatial data sets. In the figure, we can see that the execution time is linearly scalable to the database size.

Finally, please note that, we only reported results from the data set I1M.C15.R50 due to the page limit. Indeed, the results from other data sets are consistent with the above presented results.

8 Conclusions

In this paper, we formalized the problem of mining co-location patterns with rare spatial features. We first introduced a new measure called the maximal participation ratio (maxPR) and showed how this measure can be used to capture co-location patterns in spatial data sets with rare features. In addition, an algorithm called pruneMax was developed to exploit the weak monotonicity property of the maxPR measure and efficiently identify co-location patterns with rare features. Finally, our experimental results showed that the performance of the pruneMax algorithm is much better than an alternative, the min–max algorithm, which is a simple extension of the apriori-like solution [4].

This study opens several interesting directions for future research. First, in many applications, it is important to go beyond "support"-based pruning to find co-location patterns involving rare spatial features, but "infrequent" patterns. It would be interesting to examine whether we can carry this spirit to mine other kinds of patterns without "support"-based pruning. Second, the approach developed in this paper only deals with boolean spatial features. In the real world, the features can be categorical and continuous. There is a need to extend the co-location mining framework to handle continuous spatial features. Finally, if locations of spatial features change over time, it would be interesting to mine spatio-temporal association patterns.

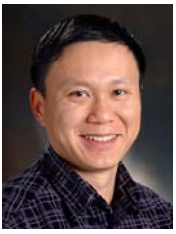
References

1. E. Brockovich, http://www.masryvititoe.com/erin_brockovich.shtml.
2. "West Nile disease," in http://www.cdc.gov/ncidod/dvbid/westnile/index.htm.
3. R. Agarwal, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases," in *Proc. of the ACM SIGMOD Conference on Management of Data*, Washington, DC, pp. 207–216, 1993.
4. 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.
5. L. Arge, O. Procopiuc, S. Ramaswamy, T. Suel, and J. Vitter. "Scalable sweeping-based spatial join," in *Proc. of the Int'l Conference on Very Large Databases*, Morgan Kaufman, San Mateo, CA, pp. 570–581, 1998.
6. Y. Chou. Exploring Spatial Analysis in Geographic Information System. Onward Press: Santa Fe, NM ISBN:1566901197, 1997.

7. N.A.C. Cressie. Statistics for Spatial Data. Wiley: New York ISBN:0471843369, 1991.
8. Environmental Systems Research Institute, Inc. "ArcGIS Family," in http://www.esri.com.
9. M. Ester, A. Frommelt, H.-P. Kriegel, and J. Sander. "Algorithms for characterization and trend detection in spatial databases," in *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, pp. 44–50, 1998.
10. 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 Geocomputation*, pp. 24–26, 2001.
11. 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.
12. R.H. Guting. "An introduction to spatial database systems," *Very Large Data Bases Journal*, Vol. 3: 357–399, 1994.
13. Y. Huang, H. Xiong, S. Shekhar, and J. Pei. "Mining confident co-location rules without a support threshold," in *Proceedings of the 18th Annual ACM Symposium on Applied Computing (SAC'03)*, Melbourne, Florida, pp. 497–418, 2003.
14. J. M. Patel, and D. J. DeWitt. "Partition based spatial-merge join," in *Proc. of the ACM SIGMOD Conference on Management of Data*, pp. 259–270, 1996.
15. E. M. Knorr, R. T. Ng, and D. L. Shilvock. "Finding boundary shape matching relationships in spatial data," in *Proc. 5th International Symposium on Spatial Databases*, Springer, Berlin Heidelberg New York, pp. 29–46, 1997.
16. K. Koperski, J. Adhikary, and J. Han. "Spatial data mining: Progress and challenges," in *Workshop on Research Issues on Data Mining and Knowledge Discovery*, pp. 409–418, Oxford University Press, UK, 1996.
17. 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.
18. M. Koubarakis, T.K. Sellis, A.U. Frank, S.Grumbach, R.H. Güting, C.S. Jensen, N.A. Lorentzos, Y. Manolopoulos, E. Nardelli, B. Pernici, H.-J. Schek, M.Scholl, B. Theodoulidis, and N. Tryfona. Spatio-Temporal Databases: The CHOROCHRONOS Approach. Springer: Berlin Heidelberg New York, 2003.
19. S.T. Leutenegger, and M.A. Lopez. "The Effect of buffering on the performance of R-trees," in *Proc. of the Int'l Conference on Data Engineering*, IEEE Educational Activities Department, pp. 164–171, 1998.
20. 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.
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. R. T. Ng, and J. Han. "Efficient and effective clustering methods for spatial data mining," in *20th International Conference on Very Large Data Bases*, Morgan Kaufman, San Mateo, CA, pp. 144–155, 1994.
23. J.F. Roddick, and M. Spiliopoulou. "A bibliography of temporal, spatial and spatio-temporal data mining research," in *ACM Special Interest Group on Knowledge Discovery in Data Mining Explorations*, New York, pp. 34–38, 1999.
24. S. Shekhar, and S. Chawla. Spatial Databases: A Tour. Prentice Hall: New Jersey ISBN: 0130174807, 2003.
25. S. Shekhar, S. Chawla, S. Ravada, A. Fetterer, X. Liu, and C.T. Lu. "Spatial databases: Accomplishments and research needs," *IEEE Trans. Knowl.Data Eng.*, Vol. 11(1):45–55, 1999.
26. 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.
27. S. Shekhar, C.T. Lu, and P. Zhang. "Detecting graph-based spatial outliers: Algorithms and applications," in *The Seventh ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, San Francisco,California, pp. 371–376, 2001.
28. S. Shekhar, P. Schrater, W.R. Raju, W. Wu, and S. Chawla. "Spatial contextual classification and prediction models for mining geospatial data," in *IEEE Transactions on Multimedia: Special Issue on Multimedia Databases*, IEEE Trans. Multimedia, pp. 174–188, 2002.
29. 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.

30. M.F. Worboys. GS: A Computing Perspective. Taylor and Francis: New York 1995.
31. 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.

Yan Huang received her B.S. degree in Computer Science from Beijing University, Beijing, China, in July 1997 and Ph.D. degree in Computer Science from University of Minnesota, Twin-cities, MN, USA, in July 2003. She is currently an assistant professor at the Computer Science and Engineering Department of University of North Texas, Denton, TX, USA. Her research interests include sensor networked databases, scientific databases, data mining, and geographic information systems (GIS). She has published over 20 technical papers in peer-reviewed journals and conference proceedings. She has served on the program committees for a number of conferences and workshops, and has been a reviwer for several journals. She is a recipient of a Ralph E. Powe Junior Faculty Enhancement Awards from ORNL Oak Ridge Associated Universities and is a member of the IEEE Computer Society, the ACM, and the ACM SIGMOD.



Jian Pei received the Ph.D. degree in Computing Science from Simon Fraser University, Canada, in 2002. He is currently an Assistant Professor of Computing Science at Simon Fraser University, Canada. His research interests can be summarized as developing effective and efficient data analysis techniques for novel data intensive applications. Particularly, he is currently interested in various techniques of data mining, data warehousing, online analytical processing, and database systems, as well as their applications in bioinformatics. His current research is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the National Science Foundation (NSF) of the United States. Since 2000, he has published over 70 research papers in refereed journals, conferences, and workshops, has served in the organization committees and the program committees of over 60 international conferences and workshops, and has been a reviewer for some leading academic journals. He is a member of the ACM, the ACM SIGMOD, the ACM SIGKDD, the IEEE Computer Society.

Hui Xiong is an assistant professor in the Management Science and Information Systems department at Rutgers, the State University of New Jersey. He received the B.E. degree in Automation from the University of Science and Technology of China, China, the M.S. degree in Computer Science from the National University of Singapore, Singapore, and the Ph.D. degree in Computer Science from the University of Minnesota, MN, USA. His research interests include data mining, statistical computing, Geographic Information Systems (GIS), Biomedical informatics, and information security. He has published over 20 technical papers in peer-reviewed journals and conference proceedings and is the co-editor of the book entitled "Clustering and Information Retrieval". He has also served on the program committees for a number of conferences and workshops. Dr. Xiong is a member of the IEEE Computer Society, the ACM, the ACM SIGKDD, and Sigma Xi.