Mining Co-location Patterns with Dominant Features

Yuan Fang, Lizhen Wang(✉), Xiaoxuan Wang, and Lihua Zhou

Department of Computer Science and Engineering,
School of Information Science and Engineering,
Yunnan University, Kunming 650091, China
{fangyuan, lzhwang, lhzhou}@ynu.edu.cn,
wangxiaoxuan1037@163.com

Abstract. The spatial co-location pattern mining discovers the subsets of spatial features which are located together frequently in geography. Most of the studies in this field use prevalence to measure a co-location pattern's popularity, namely the frequencies of a spatial feature set participating in a spatial database. However, in some cases, users are not only interested in identifying the prevalence of a feature set, but also the features playing the dominant role in a pattern. In this paper, we focus on mining dominant-feature co-location pattern (DFCP). We firstly propose a new measure, namely disparity, to measure the disparity of features in a pattern. Secondly, we formulate the DFCP mining problem to determine DFCP and extract dominant features. Thirdly, an efficient algorithm is proposed for mining DFCP. Finally, we offer an experimental evaluation of the proposed algorithms on both real data sets and synthetic data sets in terms of efficiency, mining results and significance. The results show that our method can effectively discover DFCPs.

Keywords: Dominant feature · Co-location pattern · Feature disparity

1 Introduction

Spatial co-location mining has been a problem of great practical importance due to its broad applications for environmental protection [15], public transportation [14], location-based service [11] and urban planning [12]. Most of the studies in this field adopt a Participation Index [1] to measure a co-location pattern's prevalence, namely the frequency of a spatial feature set which locate together in a spatial database. For example, {Restaurant, Supermarket, Coffee shop} is a prevalent co-location pattern which means the restaurant, supermarket, and coffee shop are located together frequently. However, in some cases, users are not only interested in identifying the prevalence of a pattern, but also the features playing the dominant role in a pattern.

Identifying the dominant features within a co-location pattern can not only provide indications of which features are co-located frequently but also help to reveal which certain features dominate the rest of features in a co-location pattern. Therefore, it is important to determine whether a co-location pattern has dominant features and extract the dominant features from prevalent co-location patterns in certain applications. One

A. Bouguettaya et al. (Eds.): WISE 2017, Part I, LNCS 10569, pp. 183–198, 2017.
DOI: 10.1007/978-3-319-68783-4_13

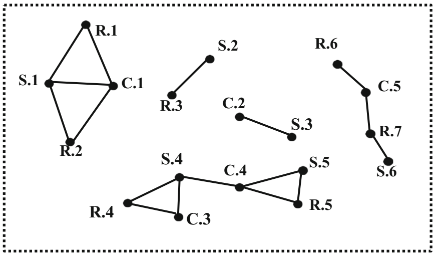


Fig. 1. Dataset of {R, S, C}

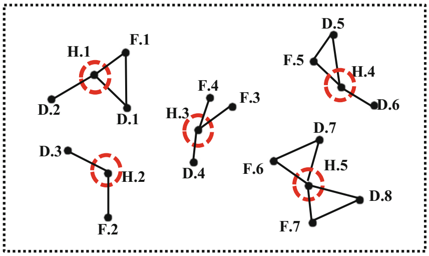


Fig. 2. Dataset of {H, D, F}

example of such applications is extracting dominant species from plant communities. Even though prevalent co-location mining on the vegetation data can discover the coexistence relation of plants, botanists require additional information on the dominant species among the prevalent co-located plant species. Another application is identifying the dominant facility in the facility points-of-interest of urban. In urban planning, the co-location information can be used to analyze the consistency between neighboring facilities. Identifying dominant facility can further support rational urban planning and business decision making. Therefore, it is necessary to identify the co-location patterns with dominant features.

For example, {Restaurant(R), Supermarket(S), Coffee shop(C)} and {Hospital(H), Drugstore(D), Flower store (F)} are two prevalent co-location patterns with the same participation index, their spatial distribution is shown in Figs. 1 and 2 respectively. The points represent the instances of features, and the lines between each two points illustrate the neighbor relationship between the two instances accordingly. From the perspective of pattern prevalence, the two co-location patterns seem to be very similar. However, in Fig. 1, if any feature of the pattern is deleted, the rest features will still be co-located frequently. This is to say that none of the features can dominate other features in Fig. 1, the status of all the features in this pattern are equivalent. Therefore, it indicates no dominant feature in the co-location pattern {Restaurant, Supermarket, Coffee shop}. On the contrary, in Fig. 2, the features “Flower store” and “Drug store” always appear simultaneously with the feature “Hospital”. More specifically, the feature “Hospital” is always the center feature clustered around by the other two features. There are many instances of “Flower store” or “Drug store” neighboring to “Hospital” separately. Nevertheless, there is no additional proximity relation between “Flower store” and “Drug store” without “Hospital”. According to this observation, it can be argued that “Hospital” is the dominant feature in the co-location pattern {Hospital, Drugstore, Flower store}. According to whether a pattern contains a dominant feature or not, different prevalent co-location patterns can be classified either as non-dominant-feature co-location patterns (Non-DFCPs) or as dominant-feature co-location patterns (DFCPs). The features within a Non-DFCP are correlated with each other while taking an equal position in the pattern. In a DFCP, some particular features dominate the rest of the features. The position of dominant features and the other features in a DFCP is nonequivalent. The aforementioned pattern {Restaurant, Supermarket, Coffee shop} is a Non-DFCP and {Hospital, Drugstore, Flower store} is a DFCP.

With above discussion as the starting point, this paper focuses on mining DFCPs from prevalent co-location patterns. The contributions of our work can be summarized as follows: (1) proposing the new concepts of dominant feature and dominant-feature co-location pattern (DFCP); (2) proposing a new measure, namely disparity, to measure the feature position in a co-location pattern; (3) formulating the problem of mining DFCP and proposing an efficient algorithm to identify the DFCPs and corresponding dominant features; (4) evaluating the proposed method with existing traditional co-location pattern mining method on both synthesized and real data sets in terms of efficiency, mining results, and significance.

The remainder of the paper is organized as follows: Sect. 1 offers an introduction on related works; Sect. 3 presents the basic concepts and describes related measures; Sect. 4 demonstrates an efficient DFCP mining algorithm; the experimental evaluation is discussed in Sect. 5; Sect. 6 ends this paper with some conclusive remarks.

2 Related Works

For over a decade, the co-location pattern mining has been attracting abundant research interests and widely used in many applications such as environmental protection [15], public transportation [14], location-based service [11] and urban planning [12], etc. Shekhar and Huang [1] first defined the concept of spatial co-location patterns. The co-location mining aims to find all subsets of spatial features, which located together frequently. In literature [1], the authors proposed to use participation index to measure the prevalence of a co-location. According to the anti-monotone property of the participation index, a Join-based co-location pattern mining algorithm was proposed. However, this original algorithm is subject to a large amount of instance join operations. Since then, various algorithms have been developed such as the Join-Less algorithm [2], the Partial-join algorithm [3], CPI-tree algorithm [4], iCPI-tree algorithm [5] and Order-Clique-Based algorithm [6] to avoid the expensive join operation and improve the efficiency. Although these developments have made significant contributions to algorithm efficiency, due to the vastness of spatial data, the typical co-location mining framework always leads to large collections of results, which make people hardly understand and identify the targeted ones. This has become one of the biggest obstacles in the studies and applications of co-location mining. In order to resolve this problem, many researchers have done many works to reduce the number of pattern results by mining the co-location pattern with a specific relationship or a specific target. Literature [6–8] focused on mining maximal co-location patterns, top-k closed co-location patterns and representative co-location patterns respectively, which effectively reduce the prevalent co-location results. In order to improve the applicability of the co-location patterns, many researchers addressed to find the co-locations with specific target based on expert knowledge such as high utility co-location pattern mining [9], ontology-based co-location pattern mining [10], co-location pattern mining under domain-constrains [13]. For discovering more interesting knowledge hidden in prevalent co-location pattern, some researchers have also committed to finding the co-location with a specific relationship such as causal relationship [17], competitive relationship [16].

Extracting the co-locations with specific relationships not only reveals the substantive connection between spatial features but also reduces the number of prevalent co-location pattern results. The dominant relation is widely appeared both in nature and in human society. There is a lot of work on dominant factor analysis in many specific applications such as public safety [18], power system [19], audio recognition [20], dominant species identification and urban planning. Therefore, mining co-location patterns with dominant relation and extracting dominant features is a significant but challengeable task.

3 Preliminary and Problem Formulation

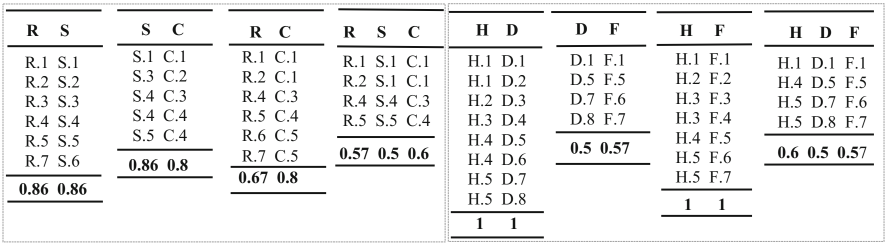
This section firstly offers a review on the preliminary concepts of typical prevalent co-location pattern mining framework, then a formal definition of feature disparity to measure the feature disparity. Furthermore, the definitions of proposed dominant feature and dominant-feature co-location pattern (DFCP) are provided and finally, the problem of DFCP mining is formulated.

3.1 Basic Concept

Given a set of spatial features $F = \{f_1, f_2 \ldots, f_n\}$, a set of their spatial instances $S = S_1 \cup S_2 \cup \ldots \cup S_n$, where $S_i(1 \leq i \leq n)$ is a set of instance of feature f_i, and a **spatial neighbor relationship** between instances R over S, the Euclidean metric is used for the R. Each instance x of feature f_i is represented as $f_i.x$, two instances $f_i.x$ and $f_j.y$ are neighbors of each other if the Euclidean distance between them is not greater than a distance threshold d. A ***k*-size co-location** $c = \{f_1, \ldots, f_k\}$ is a subset of F ($c \subseteq F$). $l = \{f_1.x_1, f_2.x_2, \ldots, f_k.x_k\}$ ($l \subseteq S$) is called a **row-instance** of c while l includes all the feature types of c and forms a **clique relationship** under R. The set of all the row-instances of c is called **table-instance** $T(c)$.

Typically, the prevalence of a k-size co-location $c = \{f_1, \ldots, f_k\}$ is measured by the **participation index** and **participation ratio**. The participation ratio $\mathrm{PR}(c,f_i)$ $(f_i \in c)$ for feature type f_i in c is the fraction of feature f_i which participates in the table instance of c. The participation ratio is defined as $\mathrm{PR}(c,f_i) = \frac{|\pi_{f_i}(T(c))|}{|T(\{f_i\})|}$, where π is the relational projection operation. The **participation index PI(*c*)** of c is the minimum participation ratio $\mathrm{PR}(c,f_i)$ in all features f_i in c: $\mathrm{PI}(c) = \min_{i=1}^{k}(\mathrm{PR}(c,f_i))(f_i \in c)$. Given a user-specified prevalence threshold *min_prev*, a co-location c is **prevalent** if PI (c) $\geq$ *min_prev*.

Example 1. Figure 2 is a spatial database with 3 features: H has 5 instances, D has 8 instances and F has 7 instances. H.1 is the first instance of feature H. Co-location {H, D, F} is the super co-location of {H, D}, {D, F} and {H, F}. The co-location instances of all the co-locations in Fig. 1 are shown in Fig. 3(a) and the co-location instances of all the co-locations in Fig. 2 are shown in Fig. 3(b) respectively. In Fig. 3 (b), suppose *min_prev* = 0.4, for co-location pattern {H, D, F}, the table instance of {H, D, F} is shown in Fig. 3(a), the PI({H, D, F}) = min(PR({H, D, F}, H), PR ({H, D, F}, D), PR({H, D, F}, F)) = min(0.6, 0.5, 0.57) = 0.5 $\geq$ *min_prev*, so c is a prevalent co-location pattern. Similarly, the PI({R, S, C}) = 0.5.



R	S
R.1	S.1
R.2	S.2
R.3	S.3
R.4	S.4
R.5	S.5
R.7	S.6
0.86	**0.86**

S	C
S.1	C.1
S.3	C.2
S.4	C.3
S.4	C.4
S.5	C.4
0.86	**0.8**

R	C
R.1	C.1
R.2	C.1
R.4	C.3
R.5	C.4
R.6	C.5
R.7	C.5
0.67	**0.8**

R	S	C
R.1	S.1	C.1
R.2	S.1	C.1
R.4	S.4	C.3
R.5	S.5	C.4
0.57	**0.5**	**0.6**

(a) Co-location instances of {R,S,C}

H	D
H.1	D.1
H.1	D.2
H.2	D.3
H.3	D.4
H.4	D.5
H.4	D.6
H.5	D.7
H.5	D.8
1	**1**

D	F
D.1	F.1
D.5	F.5
D.7	F.6
D.8	F.7
0.5	**0.57**

H	F
H.1	F.1
H.2	F.2
H.3	F.3
H.3	F.4
H.4	F.5
H.5	F.6
H.5	F.7
1	**1**

H	D	F
H.1	D.1	F.1
H.4	D.5	F.5
H.5	D.7	F.6
H.5	D.8	F.7
0.6	**0.5**	**0.57**

(b) Co-location instances of {H,D,F}

Fig. 3. The Table instances of {R, S, C} and {H, D, F}

3.2 Definitions

According to the discussion in Sect. 1, if a co-location pattern is DFCP, there will be disparity between the positions of its features. In order to test the feature position to identify DFCP and further extract the dominant features, we can calculate the disparity degree as follows: Firstly, we calculate the influence of each feature to the pattern. Secondly, we propose a measure, namely feature disparity, to measure the disparity between features. Thirdly, we define the dominant relation between features.

From Fig. 3, we can easily find a fact: the instances of c_{k-1} which appears in c_{k-1} but not appears in c_k mean these instances are not dominated by the feature $c_k - c_{k-1}$. Thus, for a k-size co-location pattern c_k, we can evaluate all its features influence from its k sub patterns.

Definition 1 (Loss Ratio). Given a k-size ($k > 2$) prevalent co-location pattern $c_k = \{f_1, f_2, \ldots, f_k\}$, and a sub pattern of c_k: $c_{k-1} = \{f_1, f_2, \ldots, f_{k-1}\}$, for feature $f_i (f_i \in c_{k-1}) \, (1 \le i \le k-1)$, the loss ratio of f_i from c_{k-1} to c_k is defined as follows:

$$\mathrm{LR}(c_k,\, c_{k-1}, f_i) = \frac{|\pi_{f_i}(T(c_{k-1}))| - |\pi_{f_i}(T(c_k))|}{|T(\{f_i\})|} \tag{1}$$

Note that $\mathrm{PR}(c_k, f_i) \le \mathrm{PR}(c_{k-1}, f_i)$ is ensured according to the anti-monotonicity of participation ratio which is referred in [1], it can be confirmed that $0 \le \mathrm{LR}(c_k, c_{k-1}, f_i) \le \mathrm{PR}(c_{k-1}, f_i)$. Loss ratio describes the fraction of feature's instances loss from a co-location to its sub pattern. We then aggregate these ratios to compute a minimum single value of loss index from c_{k-1} to c_k.

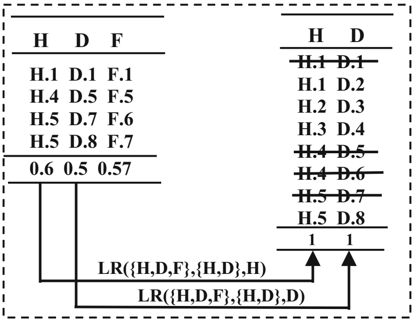
Definition 2 (Loss Index). The loss index from c_{k-1} to c_k is defined as:

$$\mathrm{LI}(c_k, c_{k-1}) = \min_{i=1}^{k-1}(\mathrm{LR}(c_k, c_{k-1}, f_i)) \tag{2}$$

Example 2. In Fig. 4, the loss ratio of feature H from {H, D, F} to {H, D} is:

$$\mathrm{LR}(\{\mathrm{H,D,F}\},\ \{\mathrm{H,D}\},\ \mathrm{H}) = \mathrm{PR}(\{\mathrm{H,D}\},\ \mathrm{H}) - \mathrm{PR}(\{\mathrm{H,D,F}\},\ \mathrm{H}) = 1 - 0.6 = 0.4$$
$$\mathrm{LR}(\{\mathrm{H,D,F}\},\ \{\mathrm{H,D}\},\ \mathrm{D}) = \mathrm{PR}(\{\mathrm{H,D}\},\ \mathrm{D}) - \mathrm{PR}(\{\mathrm{H,D,F}\},\ \mathrm{D}) = 0.5$$

The loss index LI({H, D, F}, {H, D}) = min(LR({H, D, F}, {H, D}, H), LR({H, D, F}, {H, D}, D)) = min(0.4,0.5) = 0.4. Similarly, LI({H, D, F}, {H, F}) = 0.4, LI({H, D, F}, {D, F}) = 0.



H	D	F
H.1	D.1	F.1
H.4	D.5	F.5
H.5	D.7	F.6
H.5	D.8	F.7
0.6	0.5	0.57

H	D
~~H.1~~	~~D.1~~
H.1	D.2
H.2	D.3
H.3	D.4
~~H.4~~	~~D.5~~
~~H.4~~	~~D.6~~
~~H.5~~	~~D.7~~
H.5	D.8
1	1

Fig. 4. Instance loss from {H, D} to {H, D, F}

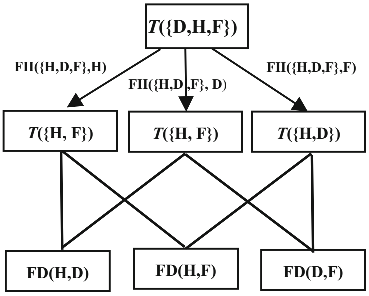


Fig. 5. The process of finding dominant features

The loss index from c_{k-1} to c_k $\mathrm{LI}(c_k, c_{k-1})$ represents the fraction of individual neighbor clique relation of c_{k-1} which is no feature $f(f = c_k - c_{k-1})$ involved, characterizes the instance loss of c_{k-1} when a new feature f join in c_{k-1} to form a higher-size pattern c_k. The higher of the loss index is, the more instances of f are irrelevant with c_{k-1}, the less dominant of feature f to the features in c_{k-1}. Therefore, we give the definition of feature influence index as follows:

Definition 3 (Feature Influence Index). Given a k-size co-location pattern c_k, the influence index of feature $f_i(f_i \in c_k)$ $(1 \leq i \leq k)$ in c_k is defined as:

$$\mathrm{FII}(c_k, f_i) = 1 - \mathrm{LI}(c_k, c_{k-1}) \tag{3}$$

where $c_{k-1} = c_k - \{f_i\}$.

Feature influence $\mathrm{FII}(c_k, f_i)$ visually represents the influence of feature f_i on other features of c_k, the higher of $\mathrm{FII}(c_k, f_i)$ is, the more instances of features in c_k may be dominated by f_i.

Example 3. In Fig. 3, for co-location pattern {H, D, F}, the feature influence of feature H in {H, D, F} is:

$$\mathrm{FII}(\{\mathrm{H,D,F}\},\mathrm{H}\}) = 1 - \mathrm{LI}(\{\mathrm{H,D,F}\},\{\mathrm{F,D}\}) = 1.$$
$$\mathrm{FII}(\{\mathrm{H,D,F}\},\mathrm{D}\}) = 1 - \ \mathrm{LI}(\{\mathrm{H,D,F}\},\{\mathrm{H,F}\})= 0.6.$$
$$\mathrm{FII}(\{\mathrm{H,D,F}\},\mathrm{F}\}) = 1 - \mathrm{LI}(\{\mathrm{H,D,F}\},\{\mathrm{H,D}\}) = 0.5.$$

Feature influence index measures the influence of single feature on other features in the same pattern. Therefore, we can use the feature influence index to measure the feature disparity. The higher the disparity between two features is, the more likely that a dominant relation exists between them.

Definition 4 (Feature Disparity). Given a k-size ($k > 2$) co-location pattern $c_k = \{f_1, f_2, \ldots, f_k\}$, the feature disparity between $f_i (f_i \in c_k)\,(1 \le i \le k)$ and $f_j (f_j \in c_k)\,(1 \le j \le k,\ i \ne j)$ in c_k is:

$$\mathrm{FD}(f_i, f_j) = |\mathrm{FII}(c_k, f_i) - \mathrm{FII}(c_k, f_j)| \tag{4}$$

According to the Definition 5, it is easily proof that feature disparity is symmetrical and non-negative.

Definition 5 (Dominant Relation). Given a k-size($k > 2$) prevalent co-location pattern $c_k = \{f_1, f_2, \ldots, f_k\}$, a minimum feature disparity threshold min_fd ($0 \le min\_fd \le 1$), the feature $f_i (f_i \in c_k)\,(1 \le i \le k)$ dominates $f_j (f_j \in c_k)\,(1 \le j \le k,\ i \ne j)$ if they meet two conditions: (1) $\mathrm{FD}(f_i, f_j) \ge min\_fd$ and (2) $\mathrm{FII}(c_k, f_i) > \mathrm{FII}(c_k, f_j)$.

It can be noticed that if there exists a dominant relation between features in a prevalent co-location, then the co-location contains dominant features.

Example 4. In Fig. 5, after calculating the loss index, we can obtain the feature influence value and then calculate the feature disparity between two features in the pattern. Setting the $min\_fd = 0.4$:

$$\mathrm{FD}(\mathrm{H}, \mathrm{D}) = \mathrm{FII}(\{\mathrm{H}, \mathrm{D}, \mathrm{F}\}, \mathrm{H}) - \mathrm{FII}(\{\mathrm{H}, \mathrm{D}, \mathrm{F}\}, \mathrm{D}) = 0.4.$$
$$\mathrm{FD}(\mathrm{H}, \mathrm{F}) = 0.5, \mathrm{FD}(\mathrm{D}, \mathrm{F}) = 0.1.$$

In {H, D, F}, FD(H, F) = 0.4 $\ge$ min_fd, FII({H, D, F}, H) > FII({H, D, F}, D), so H is a dominant feature of {H, D, F}.

Figure 5 illustrates the process for finding dominated features. We introduce two different extreme value functions – a minimum function and a maximum function – to obtain the greatest influence and least influence for a co-location pattern and further help to optimize the mining process.

Definition 6 (Max-Feature Influence Index). Given a k-size co-location pattern $c_k = \{f_1, f_2, \ldots f_k\}$, the maximum feature influence index of c_k is defined as:

$$\mathrm{max\_FII}(c_k) = \max_{i=1}^{k}(\mathrm{FII}(c_k, f_i)) \tag{5}$$

Definition 7 (Min-Feature Influence Index). Given a k-size co-location pattern $c_k = \{f_1, f_2, \ldots, f_k\}$, the minimum feature influence index of c_k is defined as:

$$\text{min\_FII}(c_k) = \min_{i=1}^{k}(\text{FII}(c_k,f_i)) \tag{6}$$

Example 5. In Fig. 3b, for co-location pattern {H, D, F}, the max-feature influence index is:

$$\begin{aligned}\text{max\_FII}(\{\text{H},\text{D},\text{F}\}) &= \max(\text{FII}(\{\text{H},\text{D},\text{F}\},\ \text{H}),\ \text{LI}(\{\text{H},\text{D},\text{F}\},\ \text{F}),\ \text{LI}(\{\text{H},\text{D},\text{F}\},\ \text{D}) \\ &= \max(1,\ 0.6,\ 0.5) = 1\end{aligned}$$

The min_feature influence index is: min_FII({H, D, F}) = 0.5

Lemma 1. If FD(f_i, f_j) $\geq$ *min_fd,* and FII (c_k,f_i) > FII (c_k,f_j), then FII (c_k,f_i) − min_FII$(c_k) \geq$ *min_fd*;

Proof. Because of FII(c_k, f_i) > FII(c_k ,f_j) $\geq$ min_FII(c_k), then

$$\text{FII}(c_k,f_i) - min\_\text{FII}(c_k) \geq \text{FD}(f_i,f_j)$$
$$\text{FII}(c_k,f_i) - \text{FII}(c_k,f_j) \geq min\_fd$$

Lemma 2. Given a k-size(k > 2) co-location pattern c_k = {f_1, f_2, ..., f_k}, two features $f_i(f_i \in c_k)\,(1 \leq i \leq k)$ and $f_j(f_j \in c_k)\,(1 \leq j \leq k)\,(i \neq j)$, if max_FII($c_k$) − min_FII ($c_k$) $\geq$ *min_fd*, c_k is a DFCP.

Proof. Assume FII(c_k, f_i) > FII(c_k, f_j), if FD(f_i, f_j) = FII(c_k, f_i) − FII(c_k, f_j) $\geq$ *min_fd,* according to Definition 5, f_i is a dominant feature, then c_k is a DFCP.

Because of FII(c_k, f_i) $\leq$ max_FII(c_k) and FII(c_k, f_j) $\geq$ min_FII(c_k), max_FII (c_k) − min_FII(c_k) $\geq$ FII(c_k, f_i) − FII(c_k, f_j) $\geq$ *min_fd*. Then, the feature which has max_FII(c_k) is the dominant feature in c_k, so c_k is a DFCP.

3.3 Problem Formulation

According Lemmas 1 and 2, we optimize the mining process. It can be formulated as follows:

Definition 8 (Max-Feature Disparity). Given a k-size (k > 2) prevalent co-location pattern $c_k = \{f_1,f_2,\ldots,f_k\}$, the max_feature disparity of the feature $f_i(f_i \in c_k)\,(1 \leq i \leq k)$ in c_k is defined as:

$$\text{max\_FD}(c_k,f_i) = |\text{FII}(c_k,f_i) - \text{min\_FII}(c_k)| \tag{7}$$

Given a minimum feature disparity threshold *min_fd* (0 $\leq$ *min_fd* $\leq$ 1), if max_FD(f_i, c_k) $\geq$ *min_fd*, *then* feature f_i is a dominant feature.

Some co-location pattern recommendation applications require that interesting patterns should be both prevalent and with dominant features. On the one hand, a dominant-feature co-location means that there is more information to support specific decision-making. Thus, it guarantees the recommendation is guiding. On the other hand, identifying the dominant-feature co-location can further decrease the number of prevalent patterns. Thus, it improves the availability of patterns. Therefore, with the definition of disparity and dominant features, we formulate the problem of mining DFCP as follows.

Mining Dominant-Feature Co-location Patterns. Given a minimum prevalence threshold *min_prev*, a minimum feature disparity threshold *min_fd*, a *k*-size co-location pattern c_k is a DFCP if it meets the following conditions:

(1) PI(c_k) $\geq$ *min_prev*
(2) max_FII(c_k) − min_FII(c_k) $\geq$ *min_fd*

Extracting Dominant Feature. Given a minimum feature disparity threshold *min_fd*, a *k*-size co-location pattern c_k, feature $f_i (f_i \in c_k)\ (1 \leq i \leq k)$ is a dominant feature in DFCP if max_FD(f_i, c_k) $\geq$ *min_fd.*

Example 6. In Fig. 3, setting prevalent threshold *min_prev* = 0.4, disparity threshold *min_fd* = 0.4. For co-location pattern {H, D, F}, PI({H, D, F})=0.5 $\geq$ *min_prev*, max_FII({H, D, F}) − min_FII({H, D, F}) = 0.5 $\geq$ *min_fd*, then {H, D, F} is a DFCP.

FII({H, D, F}, H) − min_FII({H, D, F}, H) = 1 − 0.5 $\geq$ *min_fd.*
FII({H, D, F}, D) − min_FII({H, D, F}, H) = 0.6 − 0.5 = 0.1 < *min_fd.*
FII({H, D, F}, F) − min_FII({H, D, F}, H) = 0 < *min_fd.*
H is a dominant feature in {H, D, F}.
For co-location pattern{R, S, C}: PI({R, S, C}) = 0.5 > *min_prev*
max_FII({H, D, F}) − min_FII({H, D, F}) = 0.9 − 0.71 = 0.19 < *min_fd*

We can determine that {R, S, C}is a non-DFCP prevalent co-location pattern without calculating the feature disparity between features.

4 Algorithm

In this section, we will demonstrate a general algorithm for mining DFCPs on prevalent co-location patterns. The mining framework of DFCP consists of two stages. In stage 1, the feature influence of each feature in a prevalent co-location pattern is computed, and then the DFCP is selected by a *min_fd*. Stage 2 extracts the set of dominant feature of a DFCP by a *min_fd*. Algorithm AMDFCP is the DFCP mining framework

Algorithm: AMDFCP

Input: *S*: a spatial data set; *F*:a feature set; *I*: A instances set of corresponding *F*; *d*: a distance threshold; *min_prev*: a prevalence threshold; *min_fd*: a feature disparity threshold

Output: A collection of significant co-location patterns: DFCP-set

Variables: *k*:co-location size; C_k: *k*-size candidate prevalent co-location pattern set; P_k: *k*-size prevalent co-location pattern set; PR_*c*: a collection of participation ratio of prevalent co-location pattern *c*;

Method:

```
(1)$SN$=gen_star_neighborhoods($F$,$S$,$d$);
(2)$P_1=F$, $k=2$, DFCP=$\emptyset$;
(3)WHILE($P_{k-1}\neq\emptyset$) DO
(4)$C_k$=gen_candidate_co-location($k$, $P_{k-1}$)
(5)  FOR EACH $c\in C_k$ DO
(6)     IF calculate PI($c$) $\geq min\_prev$ DO
(7)        FOR EACH  $p\in P_{k-1}(c)$ and PR_$c$ DO
(8)            LI($c,p$)= calculate_LI(PR_$p$, PR_$c$);
(9)            FII_set($c$)$\leftarrow$\{1- LI($c,p$), $c$-$p$\};
(10)       END DO
(11)       min_FII($c$)=calculate_min_FII(FII_set($c$));
(12)       max_FII($c$)=calculate_max_FII(FII_set($c$));
(13)       IF calculate max_FII($c$)-min_FII($c$)$\geq min\_fd$ DO
(14)          FOR EACH $f_i\in c$ DO
(15)              IF calculate max_FD($c$, $f_i$)$\geq min\_fd$ DO
(16)                 DFset($c$)$\leftarrow f_i$;
(17)              END DO
(18)          END DO
(19)          DFCP-set$\leftarrow$\{$c$, DFset($c$)\};
(20)       END DO
(21)    END DO
(22)  END DO
(23) $k=k+1$;
(24)END DO
```

Line 1 generates a set of star instances based on a distance threshold *d*. Lines 2–4 generate *k*-size co-location candidate pattern sets. Lines 5–22 describe DFCP mining includes: Loss index calculation, disparity test and dominant feature extraction processing. Lines 5–6 calculate participation index. Lines 7–10 compute the feature loss index and feature influence from a prevalent pattern to its sub pattern set. Lines 11–12 aggregate the feature influence to max_FII and min_FII to prune the DFCP mining and dominant feature extracting. Line 13 is a pruning strategy according to Lemma 2, which uses the full range of feature disparity, replace testing disparity of each feature pair. In line 13, if pattern *c* is DFCP, then lines 14–18 extract the dominant features through testing each feature in the pattern. Line15 is another pruning strategy, which uses the difference value between each feature influence and minimum feature influence to replace calculating the disparity between a feature and all the rest of features in a pattern. Line 19 stores the DFCPs with dominant features. Lines 3–23 are executed repeatedly and finally return a collection of DFCP set.

5 Experimental Study

In this section, comprehensive experiments are conducted to evaluate the proposed concepts and algorithms from multiple perspectives on both synthetic and real data sets. All the algorithms are implemented in Visual C#. All of our experiments are performed on a 2.4 GHZ, 8 GB-memory and Intel Core i3.

5.1 Experiment on Synthetic Data

5.1.1 Synthetic Data Generation

There are 4 synthetic datasets used in our experiments. We use different data generation methods to generate a synthetic dataset with different distributions. Dataset 1, dataset 3, and dataset 4 are randomly generated according to the Poisson distribution and distributed evenly in 1000 $\times$ 1000 space. Dataset 2 is generated for simulating the data distribution with dominant features. In order to generate the synthetic data as far as possible consistent with the dominant feature distribution characteristics, we assign different distributions of feature instances that classified as dominant features and non-dominant features respectively by controlling the specific parameters. Compared to randomly sow the data point on a plane, we increase the density of the instance of non-dominant features which around the dominant feature by setting a density parameter α and adjust the distance between the instances of same dominated feature by setting distance parameter β to ensure they are not located too closely and tightly. In our experiments, Four synthetic data sets are generated which shown in Table 1. Especially, we assign 5 dominant features and 15 non-dominant features in dataset 2.

Table 1. The experimental data set

Dataset	Instance amount	Feature amount	Data type
Dataset 1	20,000	20	Synthetic
Dataset 2	20,000	20	Synthetic
Dataset 3	40,000	25	Synthetic
Dataset 4	80,000	30	Synthetic
Dataset 5	26,546	16	Real
Dataset 6	335	32	Real

Table 2. The default values of the parameters

Parameters	Default values
Instance amount	40000
min_prev	0.3
Distance threshold	30
min_fd	0.2

We implement DFCP mining algorithm on multiple synthetic datasets and compare the efficiency of the DFCP mining algorithm and the traditional co-location Join-less [2] algorithm by considering the effect of the number of instances, the participation index threshold, the distance threshold, and the significance threshold. Table 2 shows the default parameters in the experiment.

5.1.2 Efficiency

The Effect of Prevalence Threshold. Figure 6 shows the running time of DFCP mining algorithm at five different minimum prevalence thresholds (*min_prev*) on four synthetic data sets respectively. For each data set, the run time decreases as the *min_prev* increases. For all data sets, the running time increases as the data volume increases and the *min_prev* decreases. For dataset 4, the effect of *min_prev* on algorithm performance is particularly evident, this is because low threshold and dense data lead to huge table-instances of candidates, table-instances' computation affects the performance of the algorithm. For dataset 1 and dataset 2, the running time of dataset 2 is longer than dataset 1 since the computation cost in DFCP process in dataset 2 is more than in dataset 1.

The Effect of Distance Threshold. Figure 7 depicts the running time on four synthetic data sets with respect to the variation of threshold distance. It can be observed that for each data set, the run time decreases as the distance threshold increases. For all datasets, the running time increases with the increase of the data volume and distance threshold. It can also be observed that the effect of large distance threshold on algorithm performance is especially obvious, which indicates that the performance of the algorithm is mainly affected by the data density. For dataset 1 and dataset 2, when the distance is at 10 and 20, the efficiency on dataset 1 is better than dataset 2, however, with the increase of distance, the situation has reversed. It is because when the distance is low, the computation cost of DFCP mining process affects the efficiency of dataset 2, yet when the distance is high, the auto-correlation which happened in dataset 1 is more frequent than dataset 2 so that dataset 2 performs better.

The Effect of Co-location Feature Disparity Threshold. Figure 8 demonstrates the running time on four synthetic data sets with respect to the variation of disparity threshold (*min_fd*) respectively. Comparing the running time on the four datasets, for each dataset, running time decreases more rapidly when the significance threshold increases. We note that the change of *min_fd* is particularly evident for the performance of algorithms on dense datasets. The efficiency of dataset1 is better than dataset 2 because the number of DFCP in dataset 2 is more than dataset 1 and the computation cost of dataset 2 is more than dataset 1.

5.1.3 The Mining Results of AMDFCP vs Join-Less

Figure 9 shows the mining results of AMDFCP mining algorithm and the Join-less algorithm on four synthetic data sets respectively under the default parameters. The number of DFCPs increases as the data volume increases, indicating that the denser the data, the more prominent the characteristics of the feature correlation in the prevalent co-locations. With the increase of the volume of data set, the number of DFCPs is much less than that of prevalent co-locations. Obviously, the number of DFCP in dataset 2 is far more than dataset 1, that means our algorithm can identify DFCPs correctly.

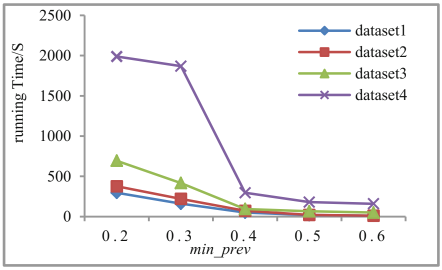


Fig. 6. Running time on different synthesized data sets (w.r.t *min_prev*)

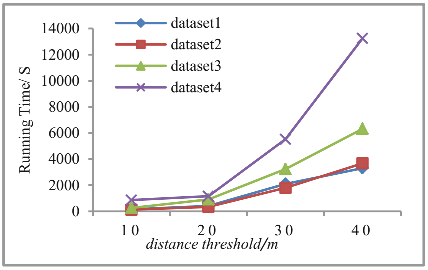


Fig. 7. Running time on different synthesized data sets (w.r.t *distance*)

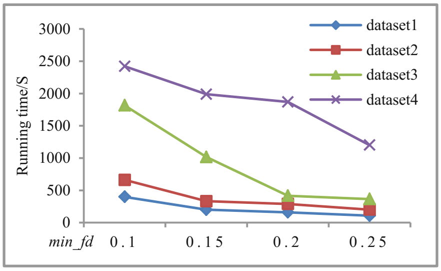


Fig. 8. Running time on different synthesized data sets (w.r.t *min_fd*)

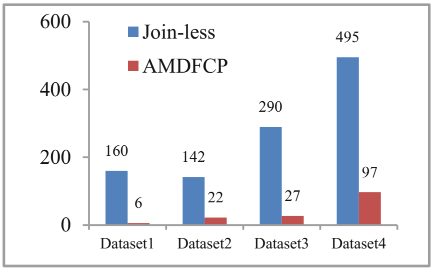


Fig. 9. The comparison of mining results on different synthesized data sets

5.2 Experiments on Real Data

Two real-world data sets are used in our experiments. The first one is the points of interest (POI) in Beijing which consists of 26,546 spatial instances and 16 spatial feature types. The spatial distance threshold is 50 by default (meaning 50 m in the real world). The second data set is the vegetation data of the “Three Parallel Rivers Area” which consists of 335 spatial instances and 32 spatial feature types. The spatial distance threshold is 6000 by default (meaning 6 km in the real world). We set the default values as *min_prev* = 0.3 and *min_fd* = 0.2.

Compared to the synthetic data, the spatial correlation of real data is higher, thus the results have practical significance. We conduct experiments to record the number of patterns for AMDFCP algorithm compared with the number of Join-less algorithms by considering the change of prevalence threshold, the distance threshold, and the feature disparity threshold. The results of varying *min_prev* on POI dataset and vegetation dataset are presented respectively in Figs. 10 and 11. It can be observed that AMDFCP algorithm reduces the number of mining results as high as 50% because with decreasing of *min_prev*, the results of the Join-less algorithm are increased quickly, AMDFCP algorithm can filter the low correlated prevalent patterns efficiently. Figures 12 and 13 illustrate the results by varying distance threshold on POI dataset and vegetation dataset respectively. The number of DFCPs increases slower than prevalent co-location patterns as the data volume increases, it indicates that the mining result is less affected by distance thresholds. This is because the disparity metrics can avoid the result explosion for aggregation of a large number of instances under the dense data.

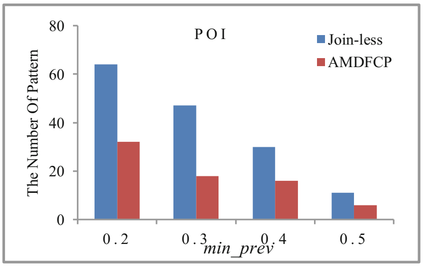


Fig. 10. The mining results with different *min_prev* on POI data set

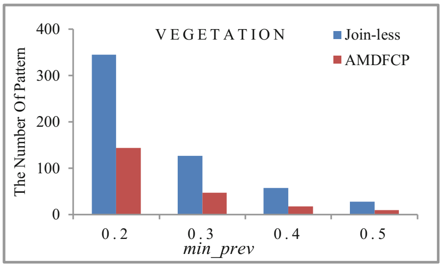


Fig. 11. The mining results with different *min_prev* on vegetation data set

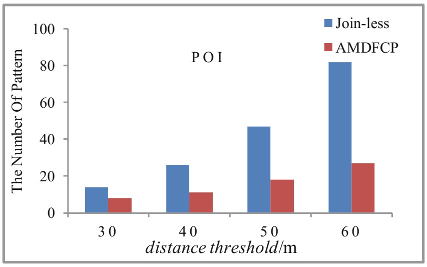


Fig. 12. The mining results with different distance threshold on POI data set

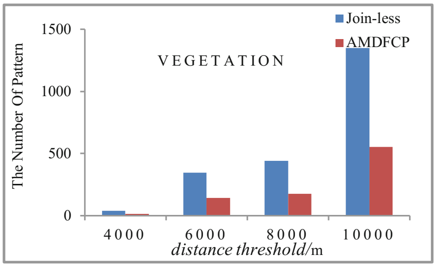


Fig. 13. The mining results with different distance threshold on vegetation data set

5.3 The Real Application of Significant Co-location Mining

We use POI datasets to present and explain the practical application of DFCP mining. Table 3 shows the results of DFCP mining, note that the dominant feature is labeled by "*". The mining results indicate that the DFCP mining can offer targeted and abundant information.

Table 3. The results of DFCP mining

DFCP	PI	FD
min_prev = 0.3; *min_fd* = 0.4; *distance threshold* = 50		
{Chinese Restaurant*, Parking Lot, clothing store}	0.42	0.4
{Chinese Restaurant, Hotel*, Parking Lot}	0.35	0.41
{Hotel*, clothing store*, Parking Lot}	0.42	0.47, 0.47
{Chinese Restaurant*, Cafe, Hotel*}	0.46	0.43, 0.4

6 Conclusions

In this study, a new approach of mining the dominant-feature co-location patterns (DFCPs) is proposed to reveal the dominant relation between features of a pattern and reduce the prevalent co-location pattern results. The DFCP mining problem consists of the DFCP determination and dominant features extracting. In order to formulate the problem of DFCP mining, we firstly propose a measure, namely disparity, to measure the feature influence disparity, then an algorithm AMDFCP is designed to mine DFCP and dominant features. Finally, experimental results demonstrated in this paper have exemplified that DFCP mining method proposed in this paper can be utilized to identify the DFCP from prevalent co-location patterns efficiently and the experimental results in POI data have also supported the proposal that DFCP mining can provide effective support to users for specific applications.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (61472346, 61662086), the Natural Science Foundation of Yunnan Province (2015FB114, 2016FA026), the Project of Innovation Team of Yunnan University.

References

1. Huang, Y., Shekhar, S., Xiong, H.: Discovering co-location patterns from spatial data sets: a general approach. TKDE **16**(12), 1472–1485 (2004)
2. Yoo, J.S., Shekhar, S.: A joinless approach for mining spatial co-location patterns. TKDE **18** (10), 1323–1337 (2006)
3. Yoo, J.S., Shekhar, S., Smith, J., Kumquat, T.P.: A partial join approach for mining co-location patterns. In: Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems (GIS 2004), pp. 241–249 (2004)
4. Wang, L., Bao, Y., Lu, J., Yip, J.: A new Join-less approach for co-location pattern mining. In: 8th IEEE International Conference on Computer and Information Technology, pp. 197–202. IEEE Press, New York (2008)
5. Wang, L., Bao, Y., Lu, Z.: Efficient discovery of spatial co-location patterns using the iCPI-tree. Open Inf. Syst. J. **3**(1), 69–80 (2009)
6. Wang, L., Zhou, L., Lu, J., Yip, J.: An order-Clique based approach for mining maximal co-locations. Inf. Sci. **179**(19), 3370–3382 (2009)
7. Yoo, J.S., Bow, M.: Mining top-k closed co-location patterns. In: IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services, pp. 100–105. IEEE Press, New York (2011)
8. Liu, B., Chen, L., Liu, C., Zhang, C., Qiu, W.: RCP mining: towards the summarization of spatial co-location patterns. In: Claramunt, C., Schneider, M., Wong, R.C.-W., Xiong, L., Loh, W.-K., Shahabi, C., Li, K.-J. (eds.) SSTD 2015. LNCS, vol. 9239, pp. 451–469. Springer, Cham (2015). doi:10.1007/978-3-319-22363-6_24
9. Wang, X., Wang, L., Lu, J., Zhou, L.: Effectively updating high utility co-location patterns in evolving spatial databases. In: Cui, B., Zhang, N., Xu, J., Lian, X., Liu, D. (eds.) WAIM 2016. LNCS, vol. 9658, pp. 67–81. Springer, Cham (2016). doi:10.1007/978-3-319-39937-9_6
10. Bao, X., Wang, L., Chen, H.: Ontology-based interactive post-mining of interesting co-location patterns. In: Li, F., Shim, K., Zheng, K., Liu, G. (eds.) APWeb 2016. LNCS, vol. 9932, pp. 406–409. Springer, Cham (2016). doi:10.1007/978-3-319-45817-5_35

11. Yu, W.: Spatial co-location pattern mining for location-based services in road networks. Expert Syst. Appl. **46**, 324–335 (2016)
12. Yu, W., Ai, T., He, Y.: Spatial co-location pattern mining of facility points-of-interest improved by network neighborhood and distance decay effects. Int. J. Geogr. Inf. Sci. **31**(2), 280–296 (2016)
13. Flouvat, F., Soc, J., Desmier, E.: Domain-driven co-location mining. GeoInformatica **19**(1), 147–183 (2015)
14. Shekhar, S., Huang, Y.: Discovering spatial co-location patterns: a summary of results. In: Jensen, C.S., Schneider, M., Seeger, B., Tsotras, V.J. (eds.) SSTD 2001. LNCS, vol. 2121, pp. 236–256. Springer, Heidelberg (2001). doi:10.1007/3-540-47724-1_13
15. Mohan, P., Shekhar, S., Shine, J.A.: A neighborhood graph based approach to regional co-location pattern discovery: a summary of results. In: 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 122–132. ACM (2011)
16. Lu, J., Wang, L., Fang, Y., Li, M.: Mining competitive pairs hidden in co-location patterns from dynamic spatial databases. In: Kim, J., Shim, K., Cao, L., Lee, J.-G., Lin, X., Moon, Y.-S. (eds.) PAKDD 2017. LNCS, vol. 10235, pp. 467–480. Springer, Cham (2017). doi:10.1007/978-3-319-57529-2_37
17. Lu, J., Wang, L., Fang, Y.: Mining causal rules hidden in spatial co-locations based on dynamic spatial databases. In: IEEE 2016 International Conference on Computer, Information and Telecommunication Systems, pp. 1–6. IEEE Press, New York (2016)
18. Muhaya, F.: Dominant factors in national information security policies. J. Comput. Sci. **6**(7), 808–812 (2010)
19. Peng, Y., Dong, H.: Dominant factors mining in power system based on clustering analysis. Electr. Power **39**(12), 16–19 (2006)
20. Gu, J., Lu, L., Cai, R., Zhang, H.-J., Yang, J.: Dominant feature vectors based audio similarity measure. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3332, pp. 890–897. Springer, Heidelberg (2004). doi:10.1007/978-3-540-30542-2_110