A Combined Co-location Pattern Mining Approach for Post-Analyzing Co-location Patterns

Yuan Fang, Lizhen Wang*, Junli Lu and Lihua Zhou

Department of Computer Science and Engineering, School of Information Science and Engineering, Yunnan University, Kunming 650091, China

*Corresponding author

Abstract—The co-location pattern mining discovers the subsets of spatial features which are located together frequently in geography. However, the huge number of the co-location mining results limit the usability of co-location patterns. Furthermore, users hardly identify and understand the interesting knowledge directly from the single co-location pattern. In this paper, we studied the problem of extracting combined co-location patterns from a large collection of prevalent co-location patterns. We first gave the definitions of atomic co-location pattern, combined co-location pattern pair and cluster; secondly, we designed a series of interesting metrics to measure the interestingness of atomic co-location patterns, combined co-location pattern pairs and clusters; thirdly, an combined co-location mining algorithm and redundant elimination strategies were proposed. The experiments evaluated the method on real data sets and synthetic data sets. The results show that our method can effectively discover combined co-location patterns.

Keywords—co-location pattern mining; combined mining; post-analysis

I. INTRODUCTION

Spatial co-location mining has been a focused research theme in spatial data mining due to its broad applications at Earth science, environmental protection, public transportation, public safety, GIS, military systems engineering, and urban planning.

The spatial co-location mining was firstly proposed by [2] aims to discover the subsets of spatial features which are located together frequently in a spatial neighborhood. In next decade, many approaches and algorithms are proposed for improving the co-location mining efficiency [3-7], applying the co-location mining technology in different data types [8-11] (e.g. spatial uncertain data) and [12-13] introducing the applications to special cases (e.g. co-location mining with rare spatial features). Example 1 shows the co-location mining process.

Example 1: Figure 1 shows an example of spatial data set. The instances of features are represented by some points. If instances are neighbors, they are connected with a line. There are 5 spatial features $F=\{A, B, C, D, E\}$. Feature A has 6 instances, B has 5 instances, C has 5 instances, D has 4 instances and E has 3 instances.

For the co-location pattern $c=\{A, B\}$, the table instance of $cPR(c, A)=5/6, PR(c, B)=4/5, PR(c)=\text{min}(PR(c, A), PR(c, B))=0.8\leq 0.33$. The mining results show in Figure I.

The co-location mining process usually produces large collections of results. In addition, most of co-location mining algorithms adopt a level-wised search method (Apriority-like) which generates numerous redundant patterns. Because of the huge numbers of prevalent co-location patterns, users hardly identify the direct and comprehensive information they interested efficiently.

Conventional co-location pattern mining focuses on discovering the relationship in spatial features, it provides limited information for the relation in co-location patterns. In some cases, single prevalent co-location pattern may offer partial even incorrect information to users. For example, there are 5 spatial features: Restaurant, Hospital, Supermarket, School and Movie Theater. The co-location pattern mining produces 20 prevalent co-location patterns such as [restaurant, hospital]. A user wants to set some vending machines in public places like schools and hospitals, and the vending machines must keep away from the supermarkets. Unfortunately, the user may feel difficult to identify the direct information from the vast results and extract useful knowledge to support decisions. In this situation, combine the co-location patterns and make further analysis to the patterns are necessary.

Combined co-location patterns are a special case of combined patterns [14-16]. Firstly, the combined co-location pattern mining is focused on the spatial prevalent co-location patterns, the differences of our work are no concepts of transactional itemsets and no contingency table in spatial database. Secondly, the conventional transactional itemsets are dependent, spatial features are correlative, the combined method of them are different. Thirdly, analyzing spatial feature...
relations and co-location pattern relations are more complicated than transaction object relations and pattern relations for additionally considering spatial data distribution. Based on above, we open up a new prospect to detect the combined co-location pattern mining among multi-features in spatial databases.

The major challenge of co-location post-analysis is not at reducing the scale of the spatial co-location mining result set, but at extracting comprehensive and interesting information. To solve the problem, we extend the conventional combined mining technique [1] to spatial co-location mining. Through combining the co-location patterns to some groupbased on user preference, measuring their interestingness and extracting the comprehensive information, we can discover and analyze the relationship of colocation patterns. The comprehensive and interesting information corresponding to a user preference can be provided directly.

This paper proposes an approach to mine combined co-location patterns, which is based on prevalent co-locations. The main contributions of our work are 1) a series of definitions referred to combined co-location patterns, including atomic co-location patterns, combined co-location pairs and combined co-location clusters. 2)A series of interestingness metrics are designed for measuring the combined co-location patterns. 3) Redundancy eliminating strategy is proposed. 4) An efficient algorithm for combined pattern mining; and the experiment results show the performance of proposed approach.

The remainder of the paper is organized as follows: Section 2 describes related definitions and proposes interestingness measures; Section 3 presents a combined co-location mining algorithm, and the experimental evaluation is presented in Section 4; Section 5 ends this paper with the summary and the future works.

II. PROBLEM STATEMENT

In this section, we firstly give the definition of atomic co-location pattern, then introduce a series related definitions to formulate the combined co-location pattern extraction problem.

A. Definition of Combined Co-location Patterns

Compared with conventional transaction data mining, spatial data is correlated. That means during the combined co-location pattern mining process, if all features are treated equally, we have to consider all the feature combinations for each prevalent co-location pattern, the computation complexity will be NP-Hard.

In this situation, we introduce user preferences that allow users choose the features they interested.In this paper, according to users’ interests, we divide the features set into two sets: interesting set and the reference set. Interesting set stands for features which are user choose, reference set consists of the features which not in the interesting set. Each co-location pattern in prevalent co-location pattern set consists of interesting features which belong to the interesting feature set and reference features belong to the reference feature set.

Definition 1 (Atomic Co-location Pattern)

Given a feature set F, an interesting feature set \( F_I \) \( (F \subseteq F) \), a reference feature set \( F_R \) \( (FR = F \setminus F_I) \), a \( k \)-size prevalent co-location pattern \( c = \{f_1, f_2, \ldots, f_k\} \), \( c \) is divided into I set and R set: \( c : I \cup R \), where \( I \subseteq F_I, R \subseteq F_R \), \( I \neq \emptyset, R \neq \emptyset, I \cap R = \emptyset \)

\( I \cup R \) is defined as an atomic co-location pattern, atomic co-location pattern is basic unit of combined co-location patterns. Note that if a co-location pattern cannot be divided into set \( I \) and set \( R \), it is not an atomic co-location pattern.

The distribution of spatial data is correlated, that means the co-location patterns which share same features may correlated. For analyzing the relationship of the co-location patterns, we combine the atomic co-location patterns into pairs/clusters. In combination process, we hope that the set \( I \) shares enough similarity and the set \( R \) is distinct to provide enough contrast and comprehensive information.

Definition 2 (Combined Co-location Pattern Pair)

Given two atomic co-location patterns \( I \cup R_1, I \cup R_2, R_1 \) and \( R_2 \) are different subsets of \( FR \), a combined co-location pattern pair \( P \) is defined as \( P = \{P_1, I \cup R_1, P_2, I \cup R_2\} \), where \( P_1 \neq \emptyset, P_2 \neq \emptyset, P_1 \cap P_2 = \emptyset \)

Definition 3 (Combined Co-location Pattern Cluster)

A combined co-location pattern cluster is a set of atomic co-location patterns which share the same I but different R. A combined co-location pattern cluster C is defined as \( C = \{P_1, I \cup R_1, P_2, I \cup R_2, \ldots, P_n, I \cup R_n\} \), where \( I \neq \emptyset, R_1 \neq \emptyset, R_2 \neq \emptyset, \ldots, R_n \neq \emptyset \)

Note that the combined co-location pattern cluster is formed by a most interesting combined co-location pattern pairs with other atomic combined co-location patterns which share same I set. Next we introduce the interesting metrics of combined co-location pairs and clusters.

B. Interestingness Measures

For the peculiarity of the co-location mining, traditional interestingness is not appropriate for measuring combined co-location pattern mining. New interestingness metrics are designed as follows for measuring the interestingness of atomic co-location patterns, combined co-location pattern pairs and clusters.

In this paper, we use the average participation ratio of the features in \( I (R) \) set for measuring \( I (R) \) contributes to atomic co-location pattern.

Definition (Interestingness of atomic co-locationPattern)

Given an atomic co-location pattern \( I \cup R \), \( k_1 \) is the size of the set \( I \subseteq F_I \), and \( k_2 \) is the size of the set \( R \subseteq F_R \). The interestingness metric of \( I \cup R \) measures the contribution of \( I \) or \( R \) to \( I \cup R \) respectively. The interestingness of \( I \) to \( I \cup R \) is designed as follows:

\[
IA_I (I \cup R) = \frac{1}{|I|} \sum_{j=1}^{k_1} \frac{|table_{instance}(I \cup R)|}{|table_{instance}(I_j)|}
\]
Co-location ICP measures the contribution of the two different reference patterns which has maximum \(IA\) co-location pattern clusters, in this paper, we choose the atomic patterns contain same as follows:

\[
\bigcup I \quad \bigcup R
\]

**Definition 5 (Co-location Coincident Ratio)**

Given two atomic co-location patterns \(I \cup R_1, I \cup R_2\), which contain same \(I\) but different \(R\), Co-location Coincident Ratio \(CCR(I \cup R_1), (I \cup R_2)\) is defined as follows:

\[
\max_{i,j=1} \left( \frac{\left| \text{table_instance}(I \cup R_1) \cap \text{table_instance}(I \cup R_2) \right|}{\text{max} \left( \left| \text{table_instance}(I \cup R_1) \right|, \left| \text{table_instance}(I \cup R_2) \right| \right)} \right)
\]

\(CCR\) contributes the correlation of spatial features, the higher \(CCR\) means the more same instances of \(I\) exist in both \(I \cup R_1\) and \(I \cup R_2\), i.e. the two atomic co-location patterns \(I \cup R_1\) and \(I \cup R_2\) share enough similarity of \(I\), combined them is meaningless without enough contrast. Considering the similarity of \(I\) and the distinction of \(R\), the interestingness of a combined co-location pattern pair is given in next definition.

**Definition 6 (Interestingness of combined co-location pattern pairs (ICP))**

Given a combined co-location pattern pair \(CP_1: I \cup R_1, CP_2: I \cup R_2\), the interestingness of \(CP_i\) is defined as

\[
\text{ICP}(CP) = IA_R(I \cup R) \times \text{ICP}(CP_1, CP_2)
\]

In this paper, ICP(\(CP\)) is also written as ICP(\(CP_1, CP_2\)). ICP measures the contribution of the two different reference feature sets to the same interesting feature set. The higher ICP is, the more interesting a combined co-location pattern pair.

Considering set \(I\) need enough similarity in the combined co-location pattern clusters, in this paper, we choose the atomic pattern which has maximum \(IA\) to be the cluster center. The interestingness of a co-location pattern cluster is the maximum interestingness of pair which consists of the cluster center pattern and the other atomic co-location pattern in the cluster.

**Definition 7 (Interestingness of combined co-location pattern cluster)**

Given a combined co-location pattern cluster \(CC\) composed of \(n\) atomic co-location patterns \(I \cup R_1, I \cup R_2, \ldots, I \cup R_n\), the cluster center which with maximum \(IA_R(I \cup R_n)\) of all the atomic co-location patterns in the cluster, the interestingness of combined co-location pattern \(CC_1, CC_2, \ldots, CC_n\) is defined as follows:

\[
\text{ICP}(CC) = \max_{j \in \{n\}} \text{ICP}(I \cup R_i, I \cup R_j)
\]

The \(\{I \cup R_j\}\) is the cluster center, we choose the maximum interestingness combined co-location pairs consist of \(\{I \cup R_j\}\) with other \(n-1\) atomic co-location patterns. The interestingness metrics of the combined co-location pattern clusters is the maximum interestingness of combined co-location pattern pairs \(\{I \cup R_1, I \cup R_2, \ldots, I \cup R_n\}\). Users can extract the most interesting combined co-location pattern pairs, meanwhile, other atomic combined co-location patterns in the cluster can provide additional information.

The combined co-location mining results of Figure 1 are show in table I.

### C. Eliminating Redundancy Strategy

In combined pattern mining process, there are 6 types of redundancy for the correlation co-location patterns. The Redundancy Strategy will be shown in algorithm.

**Table I. The Interestingness of Combined Co-location Cluster and Redundancy Elimination**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Atomic Pattern</th>
<th>IA_R</th>
<th>IA</th>
<th>CCR</th>
<th>ICP</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>(A \cup {C})</td>
<td>0.75</td>
<td>0.66</td>
<td>P1: (A \cup {C})(A \cup {D})/0.66</td>
<td>P1:0.25</td>
<td>P1:0.25</td>
</tr>
<tr>
<td></td>
<td>(A \cup {D})</td>
<td>0.75</td>
<td>0.66</td>
<td>P2: (A \cup {C})(A \cup {E})/0.2</td>
<td>P2:0.6</td>
<td>P2:0.6</td>
</tr>
<tr>
<td></td>
<td>(A \cup {E})</td>
<td>1</td>
<td>0.5</td>
<td>P4: (A \cup {D})(A \cup {E})/0.25</td>
<td>P4:0.56</td>
<td>P4:0.56</td>
</tr>
<tr>
<td></td>
<td>(A \cup {C,D})</td>
<td>0.5</td>
<td>0.31</td>
<td>P6: (A \cup {C})(A \cup {D})/0</td>
<td>P6:0.4</td>
<td>P6:0.4</td>
</tr>
<tr>
<td>C2</td>
<td>(B \cup {E})</td>
<td>0.75</td>
<td>0.4</td>
<td>P2: (B \cup {E})(B \cup {C})/0.25</td>
<td>P2:0.25</td>
<td>P2:0.25</td>
</tr>
<tr>
<td></td>
<td>(B \cup {C,D})</td>
<td>0.5</td>
<td>0.4</td>
<td>P6: (B \cup {C})(B \cup {D})/0</td>
<td>P6:0.4</td>
<td>P6:0.4</td>
</tr>
</tbody>
</table>

III. Algorithms

In this section, we will discuss a general algorithm for mining combined co-location patterns on prevalent co-location patterns.

**Algorithm**

**Input:** \(CP\) : A collection of prevalent co-locations and their table instances; \(FI\) : a user specified interesting feature set

**Output:** A collection of combined co-location patterns : CCP

**Variables:** \(AP\): A collection of atomic combined co-location pattern  
\(I\) : the interesting part of a co-location pattern (SGF)  
\(R\) : the reference part of a co-location pattern (RGF\(\setminus FI\))

**Method:**

BEGIN


The algorithm contains 3 steps.

1. **Initiate an empty CCP**
   // Generate atomic combined co-location patterns

2. **while**(CCP≠∅)do
   **for** each co-location c in CCP:
   2.1 divide c into I and R (c is represent as \(I \cup R\))
   if cannot be divided into I and R eliminate c;
   // Construct combined co-location pattern pairs and clusters
   2.2 put atomic co-location patterns \(I \cup R\) into group with same I
   // Checking CCR redundancy
   5 for each combined co-location cluster/pair
   6 Checking redundant combined co-locations pattern pairs/clusters
   7 for each combined co-location cluster/pair M
   8 if I=Ic and R=Rc, and ICC(Mc)>ICC(M)
   9 eliminate Mc from CCP as a redundancy
   10 take descending order of CCP by ICC
   Output CCP

The algorithm contains 3 steps.

1. **Generate atomic combined co-location patterns**
2. **while**(CCP≠∅) do
3. **for** each co-location c in CCP:
   1. divide c into I and R (c is represent as \(I \cup R\))
   if cannot be divided into I and R eliminate c;
4. **// Construct combined co-location pattern pairs and clusters**
5. **2.2 put atomic co-location patterns \(I \cup R\) into group with same I
6. **// Checking CCR redundancy**
7. **5 for each combined co-location cluster/pair**
8. **// Checking redundant combined co-locations pattern pairs/clusters**
9. **7 for each combined co-location cluster/pair M**
10. **8 if I=Ic and R=Rc, and ICC(Mc)>ICC(M)**
11. **9 eliminate Mc from CCP as a redundancy**
12. **10 take descending order of CCP by ICC**
13. **Output CCP**

The algorithm contains 3 steps.

1. **Initiate an empty CCP**
   // Generate atomic combined co-location patterns

2. **while**(CCP≠∅) do
   **for** each co-location c in CCP:
   2.1 divide c into I and R (c is represent as \(I \cup R\))
   if cannot be divided into I and R eliminate c;
   // Construct combined co-location pattern pairs and clusters
   2.2 put atomic co-location patterns \(I \cup R\) into group with same I
   // Checking CCR redundancy
   5 for each combined co-location cluster/pair
   6 Checking redundant combined co-locations pattern pairs/clusters
   7 for each combined co-location cluster/pair M
   8 if I=Ic and R=Rc, and ICC(Mc)>ICC(M)
   9 eliminate Mc from CCP as a redundancy
   10 take descending order of CCP by ICC
   Output CCP

The algorithm contains 3 steps.

1. **Initiate an empty CCP**
   // Generate atomic combined co-location patterns

2. **while**(CCP≠∅) do
   **for** each co-location c in CCP:
   2.1 divide c into I and R (c is represent as \(I \cup R\))
   if cannot be divided into I and R eliminate c;
   // Construct combined co-location pattern pairs and clusters
   2.2 put atomic co-location patterns \(I \cup R\) into group with same I
   // Checking CCR redundancy
   5 for each combined co-location cluster/pair
   6 Checking redundant combined co-locations pattern pairs/clusters
   7 for each combined co-location cluster/pair M
   8 if I=Ic and R=Rc, and ICC(Mc)>ICC(M)
   9 eliminate Mc from CCP as a redundancy
   10 take descending order of CCP by ICC
   Output CCP

## IV. PERFORMANCE STUDY

In this section, due to the related research of combined co-location pattern mining has not been reported, we compared our proposed algorithm with a general co-location mining algorithm: Join-less Approach[3]. We use two kinds of datasets in our experiments: real datasets and a series of synthetic datasets. All the algorithms are implemented in Visual C#. All of our experiments are performed on a 2.20GHZ, 4GB-memory, Inter PC running Windows 7.

### A. Data Sets

For evaluating the algorithm performance, we use two kinds of datasets in our experiments: real dataset and synthetic datasets. The real dataset are points of interesting of Beijing?, this dataset consist of 51,546 spatial instances with 20 distinct features.

#### TABLE II. A DESCRIPTION OF DATA SETS USED IN EXPERIMENTS

<table>
<thead>
<tr>
<th>dataset</th>
<th>Instance Amount</th>
<th>Feature amount</th>
<th>Data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset1</td>
<td>20,000</td>
<td>20</td>
<td>synthetic</td>
</tr>
<tr>
<td>Dataset2</td>
<td>30,000</td>
<td>20</td>
<td>synthetic</td>
</tr>
<tr>
<td>Dataset3</td>
<td>50,000</td>
<td>25</td>
<td>synthetic</td>
</tr>
<tr>
<td>Dataset4</td>
<td>80,000</td>
<td>28</td>
<td>synthetic</td>
</tr>
<tr>
<td>Dataset5</td>
<td>51,546</td>
<td>20</td>
<td>real</td>
</tr>
</tbody>
</table>

### B. The Mining Results of Combined Mining vs General Co-Location Pattern Mining

We compare the combined co-location clusters discovered by our method (called CCC in the following) with the results mined by the general co-location mining approach: Join-less algorithm. Figure II shows the number of the Join-less mining results and the number of combined co-location clusters and pairs discovered by CCC over the prevalent co-location patterns. The experiment is running with different participation index threshold \(min\ prev\) in Dataset 1, the neighbor relation distance threshold is 20, the interesting feature amount is 10 in CCC algorithm.

![FIGURE II THE COMPARE OF MINING RESULTS](image)

Figure II indicates that CCC algorithm can reduce the number of conventional co-location pattern results effectively.

### C. Efficiency

To compare the time efficiency of CCC and Join-less, our experiments test the running time of CCC algorithm with increasing data size is shown in Table II. Figure III shows the time efficiency of Join-less algorithm and CCC algorithm in increasing size of datasets. The participation index threshold \(min\ prev\) is 0.4, the neighbor relation distance threshold is 20, the interesting feature amount is 10 in CCC algorithm.
The experiment in Figure III shows that the running time of CCC is stable increasing in different size of datasets. To test the effect of the interesting feature amount to efficiency of CCC with increasing interesting features, our experiments test the execution time of CCC algorithm with increasing data size shown in Table I. We test the execution time of CCC in 3 interesting features, 5 interesting features and 10 interesting features in Figure IV.

The experimental results in Figure IV shows that the running time of CCC is increasing by increasing features, the major expense of CCC algorithm is compute coincident instances of interesting features, interesting features and 10 interesting features in Figure IV.

The experimental results in Figure IV shows that the running time of CCC is increasing by increasing features, the major expense of CCC algorithm is compute coincident instances of interesting features, the running time of CCC is increase rapidly by increasing feature amount.

D. The Real Applications of Combined Co-location Mining

We use real data sets to analyze the applications of combined co-location mining. Table II is the results of combined co-location mining, the mining results shows that the combined co-location pattern clusters/pairs can offer comprehensive and direct information.

In some cases, potential information is hidden in large numbers of co-location pattern mining results. The number of combined co-location pattern clusters/pairs is significantly smaller than prevalent co-location patterns, it is available to users to obtain potential information which can not know from single co-location pattern. Table III shows the 3 interesting combined co-location pattern clusters and their interestingness. In the cluster C1 we can find the three co-location patterns are related with school, the high ICC means the table instances of these patterns are low coincidence. We can analyze that the school type in these patterns are different. Generally, the nursery schools are close to residential area, universities are close to ATM, and many adult trainingschools are close to CBD. Cluster C3 also shows the different types of supermarkets are always close to distinct areas. In cluster C2 we can find that parking lots are always close to toilet, however, when parking lots are located with CBD, there are no toilet located with them, the information may helpful to public service.

<table>
<thead>
<tr>
<th>Combined co-location pattern cluster</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1 {school, residential area}</td>
<td>0.356</td>
</tr>
<tr>
<td>C1 {school, ATM}</td>
<td></td>
</tr>
<tr>
<td>C1 {school, CBD}</td>
<td></td>
</tr>
<tr>
<td>C2 {parking lots, Public toilet, Scenic}</td>
<td>0.29</td>
</tr>
<tr>
<td>C2 {Parking lots, CBD}</td>
<td></td>
</tr>
<tr>
<td>C2 {parking lots, Public toilet, Crossroad}</td>
<td></td>
</tr>
<tr>
<td>C3 {supermarket, hotel}</td>
<td>0.22</td>
</tr>
<tr>
<td>C3 {supermarket, residential area}</td>
<td></td>
</tr>
<tr>
<td>C3 {supermarket, CBD}</td>
<td></td>
</tr>
</tbody>
</table>

Single co-location pattern cannot provide comprehensive information. In fact, the majority of co-location patterns can not represent all adjacent relationship which related to their own. We can find some potential and interesting information through combining co-locations.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we studied the combined co-location patterns over prevalent co-locations. We adopt the concept of combined mining to presented a framework for efficiently mining the combined co-location patterns from prevalent co-location set to reduce the co-location mining result and provide direct, comprehensive knowledge. We propose a general combined co-location mining method and redundancy elimination strategies. Furthermore, the combined co-location mining algorithm is presented and the experiments on real datasets and synthetic datasets indicate that our algorithm can reduce the co-location mining results and give the comprehensive information. As a future work, we plan to extend combined co-location mining to other different types of data.

ACKNOWLEDGMENT

This work was supported in part by grants (No. 61472346, No. 61262069) from the National Natural Science Foundation of China, in part by a grant (No. 2015FB149, No. 2015FB114) from the Science Foundation of Yunnan Province and in part by the Seventh of Yunnan University Graduate Student Scientific Research Project under Grant No. YNUY201444.

REFERENCES


